Nonparametric Smoothing Estimates of Time-Varying Coefficient Models With Longitudinal Data

By DONALD R. HOOVER
Departments of Epidemiology and Biostatistics,
The Johns Hopkins University, Baltimore, Maryland 21205, U.S.A.

JOHN A. RICE
Department of Statistics,
University of California, Berkeley, California 94720, U.S.A.

COLIN O. WU
Department of Mathematical Sciences,
The Johns Hopkins University, Baltimore, Maryland 21218, U.S.A.

AND LI-PING YANG
Department of Epidemiology,
The Johns Hopkins University, Baltimore, Maryland 21205, U.S.A.

April 11, 1996

Abstract

This paper considers estimation of nonparametric components in a varying-coefficient model with repeated measurements (Y_{ij}, X_{ij}, t_{ij}), $i = 1, ..., n$ and $j = 1, ..., n_i$, where $X_{ij} = (X_{ij0}, ..., X_{ijk})^T$ and (Y_{ij}, X_{ij}, t_{ij}) denote the jth outcome, covariate and time design points, respectively, of the ith subject. The model considered here is $Y_{ij} = X_{ij}^T \beta(t_{ij}) + \epsilon_i(t_{ij})$ where $\beta(t) = (\beta_0(t), ..., \beta_k(t))^T$, $k \geq 0$, are smooth nonparametric functions of interest and $\epsilon_i(t)$ is a mean zero stochastic process. The measurements are assumed to be independent for different subjects but can be correlated at different time points within each subject. Three nonparametric estimates, namely kernel, smoothing spline and locally weighted polynomial, of $\beta(t)$ are derived for such repeatedly measured data. A cross-validation criterion is proposed for the selection of the corresponding smoothing parameters. Asymptotic properties, such as consistency, rates of convergence and asymptotic mean squared errors, are established for the kernel estimates. These asymptotic results give useful insights into the reliability of our general estimation methods. An example of predicting the growth of children born to HIV infected mothers based on gender, HIV status and maternal vitamin A levels shows that

Key words and phrases: Varying-coefficient models; Nonparametric estimation; Mean squared error; Rates of convergence; Smoothing parameter; Longitudinal data.

Correspondence to: Colin O. Wu, Department of Mathematical Sciences, The Johns Hopkins University, Baltimore, MD 21218-4689

this model and the corresponding nonparametric estimates are useful in epidemiological studies.

1 Introduction

Longitudinal data occur frequently in medical and epidemiological studies where both the outcome and the covariates of a set of randomly selected subjects are repeatedly recorded on the same individuals over time. Let t be the time a measurement is recorded, $Y(t)$ and $X(t)$ be the real valued outcome of interest and the R^{k+1}, $k \geq 0$, valued covariate, respectively, observed at time t. Suppose there are n subjects, and, for each subject i, there are $n_i \geq 1$ repeated measurements of $(Y(t), X(t), t)$ over time. The jth observation of $(Y(t), X(t), t)$ of the ith subject is denoted by (Y_{ij}, X_{ij}, t_{ij}), $i = 1, ..., n$ and $j = 1, ..., n_i$, where $X_{ij} \in R^{k+1}$ is given by the column vector $X_{ij} = (X_{ij0}, ..., X_{ijk})^T$.

Under the classical linear model framework, theory and methods of regression with repeated observations have been extensively studied in the literature. These results include Pantula and Pollock (1985), Ware (1985), Jones (1987), Diggle (1988), Jones and Ackerson (1990), Jones and Boadi-Boteng (1991), among others. A summary of different types of parametric approaches can be found in Diggle, Liang and Zeger (1994). While parametric approaches are useful, questions will always arise about the adequacy of the model assumptions and the potential impact of model misspecifications on the analysis. This motivates the use of nonparametric approaches.

For nonparametric models with fixed design time points, Hart and Wehrly (1986), Altman (1990) and Hart (1991) considered kernel methods for estimating the expectation, $E(Y(t))$, without the presence of the covariate $X(t)$, and derived a class of generalized cross-validation bandwidth selection procedures. As an alternative to kernel methods, Rice and Silverman (1991) considered a class of smoothing splines and proposed a method of choosing the smoothing parameters by cross-validation in which subjects were left out one at a time. Although the existing kernel and spline methods are successful in predicting the mean change of $Y(t)$ over time, they only consider the effect of t and do not take account of other possibly important covariates.

To quantify the influence of covariates, Zeger and Diggle (1994) and Moyeed and Diggle (1994) studied a semiparametric model

$$Y_{ij} = \mu(t_{ij}) + X_{ij}^T \beta + \epsilon_i(t_{ij}) \tag{1.1}$$

where $\beta = (\beta_1, ..., \beta_k)^T$ is a vector of unknown constants in R^k, $\mu(t)$ is an arbitrary smooth function of t on the real line, and the error term $\epsilon_i(t)$ is a mean zero stochastic process.

By generalizing the methods of Hastie and Tibshriani (1990), Zeger and Diggle (1994) and Moyeed and Diggle (1994) suggested a backfitting procedure which initially estimates $\mu(t)$ by a class of kernel estimates and then iteratively estimates β and $\mu(t)$. Their results showed that this backfitting procedure had good asymptotic properties, such as consistency and desirable rates of convergence, and was useful in predicting the depletion of CD4 cells over time among HIV infected persons.

Because of the time dependent nature of the longitudinal studies, we consider in this paper a direct generalization of the model (1.1) that allows the coefficients to vary over time

$$Y_{ij} = X_{ij}^T \beta(t_{ij}) + \epsilon_i(t_{ij}) \tag{1.2}$$

where, for all $t \in R$, $\beta(t) = (\beta_0(t), ..., \beta_k(t))^T$, $k \geq 0$, are arbitrary smooth functions of t, $\epsilon_i(t)$ is a mean zero stochastic process, and X_{ij} and ϵ_i are independent. (We note that the process $\epsilon_i(t)$ need not have zero mean for each subject.) We make no assumptions on the structure of $\epsilon_i(t)$, such as it being stationary or autoregressive, for example. When the data are obtained from the cross-sectional i.i.d. sampling, (1.2) reduces to the varying-coefficient models studied by Hastie and Tibshriani (1993).

In Section 2, we present the three, computationally straightforward, nonparametric estimates (kernels, smoothing splines and locally weighted polynomials) of $\beta(t)$, and set forth cross-validation criterion for selecting the smoothing parameters. Section 3 applies model (1.2) and our estimates to an epidemiological example, for predicting growth of children born to HIV infected mothers, based on maternal vitamin A levels and children's gender and HIV status. Section 4 gives the asymptotic properties of the kernel estimates. In Section 5, we discuss some potentially useful generalizations of (1.2) and other related estimation methods. Finally, the proofs of the main results are deferred to Section 6.

2 Estimation by Linear Smoothing

Theory and applications of estimates based on kernel, spline and locally weighted polynomial methods have been extensively studied in the literature for nonparametric curve estimation with independent cross-sectional data. With properly selected smoothing parameters, these estimation methods have good asymptotic properties such as optimal rates of convergence, and usually give reliable results in real applications. Thus it is natural to extend these methods to the estimation of $\beta(t)$ for observations from longitudinal studies. Here we first give an alternative representation of $\beta(t)$, and then present our nonparametric estimates. This alternative representation provides insight into the nature of the estimates and is useful in deriving the asymptotic properties (cf. Section 4). At the end of this section, we give a description of a cross-validation criterion which for the selection of smoothing parameters.

2.1 Preliminary

To motivate the construction of our linear estimates and the analyses of their statistical properties, it is convenient to represent $\beta(t)$ in terms of the expectations of $Y(t)$ and $X(t)$.

Let $X(t)$ be a $k+1$ column vector and $\epsilon(t)$ be a mean zero stochastic process. Then $(Y(t), X(t))$ satisfies the varying-coefficient model (1.2) if

$$Y(t) = X^T(t)\,\beta(t) + \epsilon(t) \tag{2.1}$$

where $X(t)$ and $\epsilon(t)$ are independent. Multiplying both sides of (2.1) by $X(t)$ and rearranging the terms, we have

$$X(t)\,X^T(t)\,\beta(t) = X(t)\,Y(t) - X(t)\,\epsilon(t).$$

Taking expectations on both sides of the above equation, we have

$$E\left(X(t)\,X^T(t)\right)\beta(t) = E\left(X(t)\,Y(t)\right). \tag{2.2}$$

If $E\left(X(t)\,X^T(t)\right)$ is invertible, then $\beta(t)$ is unique and given by

$$\beta(t) = E\left(X(t)\,X^T(t)\right)^{-1} E\left(X(t)\,Y(t)\right). \tag{2.3}$$

If $E\left(X(t)\,X^T(t)\right)$ does not have a unique inverse, then $\beta(t)$ is not unique and the model (2.1) becomes unidentifiable. We assume for the rest of the paper that $E\left(X(t)\,X^T(t)\right)^{-1}$ exists. It is easy to show that, if $E\left(X(t)\,X^T(t)\right)$ is invertible, $\beta(t)$ as given in (2.3) uniquely minimizes the second moment $E\left[\left(Y(t) - X^T(t)\,\beta(t)\right)^2\right]$ for any given $t \in R$.

2.2 Kernel Estimates

This class of estimates is developed based on finding the unique $\beta(t) = (\beta_0(t), ..., \beta_k(t))^T$ which minimizes the locally weighted least squares criterion

$$\ell_N(t) = \sum_{i=1}^{n}\sum_{j=1}^{n_i}\left[Y_{ij} - \left(\sum_{l=0}^{k} X_{ijl}\beta_l(t)\right)\right]^2 K\left(\frac{t - t_{ij}}{h}\right)$$

where $N = \sum_{i=1}^{n} n_i$ is the total number of observations, h is a positive bandwidth which might depend on N, and $K(\cdot)$ is a Borel measurable kernel function mapping R onto R.

Let Y_i and X_i be the outcome vector and design matrix of the ith subject: $Y_i = (Y_{i1}, ..., Y_{in_i})^T$ and

$$X_i = \begin{pmatrix} X_{i10} & X_{i11} & \cdots & X_{i1k} \\ \cdots & \cdots & \cdots & \cdots \\ X_{in_i0} & X_{in_i1} & \cdots & X_{in_ik} \end{pmatrix}.$$

Let $K_i(t)$ be the diagonal matrix,

$$K_i(t) = diag\left(K\left[(t - t_{i1})h^{-1}\right], ..., K\left[(t - t_{in_i})h^{-1}\right]\right).$$

It is convenient to rewrite $\ell_N(t)$ into the following matrix form

$$\ell_N(t) = \sum_{i=1}^{n} (Y_i - X_i\beta(t))^T K_i(t)(Y_i - X_i\beta(t)). \tag{2.4}$$

For each given $t \in R$, $\beta(t)$ minimizes $\ell_N(t)$ if it satisfies the $k+1$ equations $\partial\ell_N/\partial\beta_l(t) = 0$ for all $l = 0, ..., k$. By (2.4), these equations are equivalent to

$$\left(\sum_{i=1}^{n} X_i^T K_i(t) X_i\right)\beta(t) = \sum_{i=1}^{n} X_i^T K_i(t) Y_i. \tag{2.5}$$

If $\sum_{i=1}^{n} X_i^T K_i(t) X_i$ is invertible, then (2.5) has a unique solution $\hat{\beta}(t) = \left(\hat{\beta}_0(t), ..., \hat{\beta}_k(t)\right)^T$ such that

$$\hat{\beta}(t) = \left(\sum_{i=1}^{n} X_i^T K_i(t) X_i\right)^{-1}\left(\sum_{i=1}^{n} X_i^T K_i(t) Y_i\right). \tag{2.6}$$

We note that (2.4) is a $(k+1) \times (k+1)$ system that must be solved for each t.

The estimate $\hat{\beta}(t)$ depends on the choices of the bandwidth and the kernel function. It is well known in estimation with independent cross-sectional data that the selection of bandwidth is more important than the selection of the kernel function. We will see in Section 3 and Section 4 that the selection of h also plays a crucial role in the properties of $\hat{\beta}(t)$ in the current longitudinal setting.

Remark 2.1 Higher order kernels with negative lobes are necessary to achieve certain asymptotically optimality properties in some cases (Härdle, 1990). (A d-th order kernel satisfies $\int u^j K(u)du = 0$, $j = 1, 2, \ldots, d-1$ and $\int u^d K(u)du \neq 0$). Simulation results [cf. Marron and Wand (1992)] with independent cross-sectional data have shown that the desired asymptotic properties are only effective when the sample sizes are unusually large. Thus, despite the theoretical advantages, there is frequently little motivation to use kernels with negative lobes in applications.

2.3 Smoothing Splines

Splines are piece-wise polynomials which are joined smoothly at knots. Smoothing splines are splines that minimize a particular penalized least squares criterion. Statistical properties and practical implementation of spline methods can be found in Eubank (1988) among others. Suppose that the functions $\beta_0(t), ..., \beta_k(t)$ of (1.2) are twice continuously differentiable and their second derivatives $\beta_0(t)'', ..., \beta_k(t)''$ are bounded and square integrable. A smoothing spline estimate of $\beta_0(t), ..., \beta_k(t)$ minimizes the penalized least squares criterion

$$J(\beta, \lambda) = \sum_{i=1}^{n} \sum_{j=1}^{n_i} \left\{ Y_{ij} - \left[\sum_{l=0}^{k} X_{ijl} \beta_l (t_{ij}) \right] \right\}^2 + \sum_{l=0}^{k} \lambda_l \int \left[\beta_l'' (t) \right]^2 dt \tag{2.7}$$

where $\lambda = (\lambda_0, ..., \lambda_k)^T$ are positive valued smoothing parameters which penalize the roughness of $\beta_0, ..., \beta_k$. As in univariate smoothing (Eubank, 1988), it can be shown that the minimizers of (2.7) are natural cubic splines with knots located at the distinct values of t_{ij}.

For minimizing $J(\beta, \lambda)$ of (2.7), it is convenient to represent $\beta_0, ..., \beta_k$ in terms of spline basis functions such as B-splines with knots as above. We express each β_l in the form

$$\beta_l(t) = \sum_{r=1}^{d} \gamma_{rl} B_r(t) = B^T(t) \gamma_l \tag{2.8}$$

where $d \geq 1$, $-\infty < t < \infty$, $\gamma_l = (\gamma_{1l}, ..., \gamma_{dl})^T$ are real valued coefficients, and $B(t) = (B_1(t), ..., B_d(t))^T$ is a set of basis functions. We can then find the coefficient vectors γ_l, $l = 0, 1, \ldots, k$ which minimize the quadratic functional $J(\beta, \lambda)$. For each subject i, let $X_{i \cdot l}$ be the diagonal matrix:

$$X_{i \cdot l} = diag(X_{i1l}, ..., X_{in_il})$$

Let $t_i = (t_{i1}, ..., t_{in_i})^T$, let $\beta_l(t_i) = (\beta_l(t_{i1}), ..., \beta_l(t_{in_i}))^T$, and let

$$B_i = \begin{pmatrix} B_1(t_{i1}) & \cdots & B_d(t_{i1}) \\ \cdots & \cdots & \cdots \\ B_1(t_{in_i}) & \cdots & B_d(t_{in_i}) \end{pmatrix}.$$

Then the first term at the right hand side of (2.7) can be written as

$$\left(Y_i - \sum_{l=0}^{k} X_{i \cdot l} B_i \gamma_l \right)^T \left(Y_i - \sum_{l=0}^{k} X_{i \cdot l} B_i \gamma_l \right).$$

Furthermore, let Ω be the $d \times d$ matrix whose (i, j)th element is given by

$$\Omega_{ij} = \int B_i''(t) B_j''(t) dt.$$

The second term at the right hand side of (2.7) is given by $\sum_{l=0}^{k} \lambda_l \gamma_l^T \Omega \gamma_l$. Thus, (2.7) is equivalent to

$$J(\beta, \lambda) = \sum_{i=1}^{n} \left(Y_i - \sum_{l=0}^{k} X_{i \cdot l} B_i \gamma_l \right)^T \left(Y_i - \sum_{l=0}^{k} X_{i \cdot l} B_i \gamma_l \right) + \sum_{l=0}^{k} \lambda_l \gamma_l^T \Omega \gamma_l. \tag{2.9}$$

Setting each $\partial J(\beta, \lambda) / \partial \gamma_{rl} = 0$, the minimizer $(\gamma_0, ..., \gamma_k)$ of (2.9) satisfies the normal equations

$$\sum_{i=1}^{n} \left[(X_{i \cdot l} B_i)^T \sum_{l=0}^{k} X_{i \cdot l} B_i \gamma_l \right] + \sum_{l=0}^{k} \lambda_l \Omega \gamma_l = \sum_{i=1}^{n} (X_{i \cdot l} B_i)^T Y_i, \quad l = 0, ..., k. \tag{2.10}$$

Let $M_{il} = X_{i \cdot l} B_i$. Then, after rearranging the terms, (2.10) is equivalent to the equations

$$\sum_{j=0}^{k} \left[\left(\sum_{i=1}^{n} M_{il}^T M_{il} \right) + \lambda_j \Omega \right] \gamma_j = \sum_{i=1}^{n} M_{il}^T Y_i, \quad l = 0, ..., k. \tag{2.11}$$

If the normal equations (2.10) have a unique solution $(\hat{\gamma}_0, ..., \hat{\gamma}_k)$, without loss of generality there exist $d \times n_i$ matrices N_{il}, for $i = 1, ..., n$ and $l = 0, ..., k$, so that

$$\hat{\gamma}_l = \sum_{i=1}^{n} N_{il} Y_i, \quad l = 0, ..., k. \tag{2.12}$$

The corresponding linear estimates $\hat{\beta}_0(t)$, ..., $\hat{\beta}_k(t)$ are obtained by substituting $(\gamma_0, ..., \gamma_k)$ with $(\hat{\gamma}_0, ..., \hat{\gamma}_k)$ in (2.8), i.e.

$$\hat{\beta}_l(t) = \sum_{r=1}^{d} \hat{\gamma}_{rl} B_r(t) = \sum_{i=1}^{n} B^T(t) N_{il} Y_i, \quad l = 0, ..., k. \tag{2.13}$$

Here the existence and uniqueness of the solution $(\hat{\gamma}_0, ..., \hat{\gamma}_k)$ of this linear system depend on the design matrices X_1, ..., X_n, t_i for $i = 1, ..., n$, and the basis functions $B_1(\cdot)$, ..., $B_d(\cdot)$. For practical implementation of smoothing splines, one has to select an adequate smoothing parameter vector λ and basis functions. The role of λ is similar to that of h in kernel estimates: a proper choice of λ will result optimal convergence rates of $\hat{\beta}_0(t)$, ..., $\hat{\beta}_k(t)$. It can be seen from (2.7) that too large a λ_l gives an excessive penality for the roughness of β_l, thus results in an over-smoothed estimate $\hat{\beta}_l$. Conversely, too small a λ_l results in an under-smoothed $\hat{\beta}_l$.

In practice, if the unique solution of (2.10) exists, it may be found directly or by using the "backfitting algorithm" suggested by Hastie and Tibshriani (1993). We note that (2.10) comprises a system of equations of order $(k+1)d \times (k+1)d$, the solution of which can be used to find the estimates for all t. In comparison with the kernel method in which a small system must be solved for each t, here a large system must be solved once. A practical difficulty is that d can be quite large since, as noted above, it is of the order of the number of distinct t_{ij}. Backfitting is one way of coping with this difficulty. It can also be quite adequately circumvented by approximating the smoothing spline solutions by splines with a relatively small number of fixed equispaced knots as in Parker and Rice (1984), thus drastically reducing the dimensionality of the computations.

Clearly, deeper theoretical properties of $\hat{\beta}_0(t)$, ..., $\hat{\beta}_k(t)$ with longitudinal observations deserve further investigation, but are beyond the scope of this paper.

2.4 Locally Weighted Polynomials

This class of estimates is a generalization of the kernel type estimates, for which theory and applications with independent cross-sectional data have been studied by Stone (1977), Cleveland (1979), Buja, Hastie and Tibshriani (1989), Hastie and Tibshriani (1990), Fan (1993) among others. This generalization has many advantages over the kernel methods, particularly in estimation at boundary points [cf. Hastie and Loader (1993)]. Theoretical and simulation results indicate that smoothings with locally weighted polynomials are effective alternatives to the smoothing splines.

Motivated by their performance in cross-sectional data, we propose a class of smoothing methods which extends the existing approaches of locally weighted polynomials to longitudinal data. Let $W_{ij}(t)$, $i = 1, ..., n$, $j = 1, ..., n_i$, be weight functions of t_{ij} and t. In particular, $W_{ij}(t)$ may be selected as a kernel function $K((t - t_{ij})/h)$ as it is used in (2.4), or based on nearest neighbors as in Section 2.11 of Hastie and Tibshriani (1990). The $(d-1)$-Degree Locally Weighted Polynomial Estimate (LWPE) of $\beta_l(t)$, $l = 0, ..., k$, is given by

$$\hat{\beta}_l(t) = \sum_{r=1}^{d} \hat{b}_{rl}(t)\, t^{r-1} = \mathcal{B}^T(t)\, \hat{b}_l(t) \tag{2.14}$$

where $\mathcal{B}(t) = \left(1, t, ..., t^{d-1}\right)^T$ are basis functions of the $(d-1)$-degree polynomial and the coefficients $\hat{b}_l(t) = \left(\hat{b}_{1l}(t), ..., \hat{b}_{dl}(t)\right)^T$ minimize the locally weighted sum of squares

$$L_N(t) = \sum_{i=1}^{n} \sum_{j=1}^{n_i} \left(Y_{ij} - \sum_{l=0}^{k} \left[X_{ijl} \mathcal{B}^T(t_{ij})\, b_l(t) \right] \right)^2 W_{ij}(t) \tag{2.15}$$

with $b_l(t) = (b_{1l}(t), ..., b_{dl}(t))^T$ and $b_{rl}(t)$ being real valued functions of t.

Let $\mathcal{B}_i$ and $\mathcal{W}_i$ be the basis matrix and the diagonal weight matrix of the ith subject such that

$$\mathcal{B}_i = \begin{pmatrix} 1 & t_{i1} & \cdots & t_{i1}^{d-1} \\ 1 & t_{i2} & \cdots & t_{i2}^{d-1} \\ \cdots & \cdots & \cdots & \cdots \\ 1 & t_{in_i} & \cdots & t_{in_i}^{d-1} \end{pmatrix} \quad \text{and} \quad \mathcal{W}_i = diag\left(W_{i1}(t), ..., W_{in_i}(t)\right).$$

The equivalent matrix form of (2.15) is

$$L_N(t) = \sum_{i=1}^{n} \left(Y_i - \sum_{l=0}^{k} X_{i \cdot l} \mathcal{B}_i b_l(t)\right)^T \mathcal{W}_i(t) \left(Y_i - \sum_{l=0}^{k} X_{i \cdot l} \mathcal{B}_i b_l(t)\right). \tag{2.16}$$

Denote

$$\frac{\partial L_N}{\partial b_l}(t) = \left(\frac{\partial L_N}{\partial b_{1l}}(t), ..., \frac{\partial L_N}{\partial b_{dl}}(t)\right)^T, \quad l = 0, ..., k,$$

and set $\partial L_N / \partial b_l(t) = 0$, for all $l = 0, ..., k$. If $\left(\hat{b}_0(t), ..., \hat{b}_k(t)\right)$ is a unique minimizer of (2.15), it satisfies the generalized normal equations

$$\sum_{i=1}^{n} (X_{i \cdot l} \mathcal{B}_i)^T \mathcal{W}_i \left(Y_i - \sum_{r=0}^{k} X_{i \cdot r} \mathcal{B}_i b_r(t)\right) = 0, \quad l = 0, ..., k. \tag{2.17}$$

To simplify the notation, let

$$\mathcal{M}_{il}(t) = (X_{i \cdot l} \mathcal{B}_i)^T \mathcal{W}_i(t), \quad \mathcal{N}_{rl}(t) = \sum_{i=1}^{n} (X_{i \cdot l} \mathcal{B}_i)^T \mathcal{W}_i(t) (X_{i \cdot r} \mathcal{B}_i),$$

$b(t) = (b_0(t), ..., b_k(t))^T$ and $\mathcal{N}(t)$ be the matrix whose (r, l)th block is $\mathcal{N}_{rl}(t)$ with $r = 0, ..., k$ and $l = 0, ..., k$. Then (2.17) is equivalent to

$$\mathcal{N}(t)\ b(t) = \mathcal{M} \circ Y \tag{2.18}$$

where $\mathcal{M} \circ Y$ is the $d + k + 1$ column matrix $\left(\sum_{i=1}^{n} \mathcal{M}_{i0} Y_i, ..., \sum_{i=1}^{n} \mathcal{M}_{ik} Y_i\right)^T$.

If $\mathcal{N}(t)$ is invertible, then (2.18) implies

$$\hat{b}(t) = \mathcal{N}^{-1}(t) (\mathcal{M} \circ Y). \tag{2.19}$$

Thus, $\hat{\beta}(t) = \left(\hat{\beta}_0(t), ..., \hat{\beta}_k(t)\right)^T$ is uniquely defined by substituting $\hat{b}(t)$ of (2.19) into (2.14). In order to compute the estimate a system of size $d(k+1) \times d(k+1)$ must be solved for each t.

Remark 2.2. Comparing (2.14) with the kernel estimates as defined in (2.6), it is easy to see that (2.6) is a special case of (2.14) with $d = 1$ and $W_{ij}(t)$ being selected as a kernel function. Here there are two smoothing parameters for the LWPEs of (2.14). One of them is the window size which is incorporated into the weight functions. When $W_{ij}(t)$ is a kernel function, the bandwidth is the corresponding smoothing parameter for window size. When $W_{ij}(t)$ is the weight function defined on the k nearest-neighbors (cf. Section 2.11 of Hastie and Tibshriani (1990)), then the size of the neighborhood k is the window size. Another parameter is the degree of the local polynomial. Of these two parameters, the selection of the window size is usually more important and may affect the rates of convergence for the estimates. By selecting a different local polynomial, one may improve the constants of the asymptotic mean squared errors of the estimates and their statistical properties near the boundary. The usual choices here are local linear or quadratic polynomials. ♠

Remark 2.3. Since LWPEs are constructed based on direct generalizations of the kernel estimates, they are equally intuitive and have known better statistical properties than kernel methods. Computationally, all three methods, i.e. kernel estimates, smoothing splines and LWPEs, require solving some linear systems and hence are comparable in terms of computation time. However, further theoretical and simulation based comparisons between these different smoothing methods are warranted. ♠

2.5 Selection of Smoothing Parameters

In practical implementation of the above linear smoothing methods, the smoothing parameters can be selected subjectively by examining scatter plots and the fitted curves. However, especially when more than one covariate is present, it is useful to develop automatic procedures so that an adequate smoothing parameter can be directly suggested by the data.

Here we will focus on a method of cross-validation for smoothing parameter choice. Following Rice and Silverman (1991), we use a form of cross-validation in which single subjects are deleted one at a time, rather than single responses, since the latter procedure is unsuitable when there is intra-subject correlation. Although theoretical properties of this cross-validation method are still yet to be developed, the main advantage of the method is that it does not rely on specific correlation structures of the data. An alternative, which we have not pursued, would be modifications for longitudinal data of plug-in procedures which have been developed independent cross-sectional data (Hall et al. (1991) and Fan and Marron (1992), for example).

Smoothing in the current context could have a number of different objectives, especially when $\beta(t)$ has more than one component. For notational simplicity, we denote by λ the smoothing parameters of any linear estimates of this section. Interest might focus on the risk of a single component $\beta(t)$ either at a fixed t or globally as measured by mean squared error

$$\text{MSE}_\lambda\left(\hat{\beta}_l(t)\right) = E\left[\left(\hat{\beta}_l(t) - \beta_l(t)\right)^2\right], \quad l = 0, ..., k. \tag{2.20}$$

or average mean squared error (AMSE)

$$\text{AMSE}_\lambda\left(\hat{\beta}_l\right) = \frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{n_i} E\left[\left(\hat{\beta}_l(t_{ij}) - \beta_l(t_{ij})\right)^2\right], \quad l = 0, ..., k. \tag{2.21}$$

Alternatively, one could focus on the local fitness of all components of $\hat{\beta}(t)$ at any given values (x, t) as measured by

$$\text{MSE}_\lambda\left(x^T\hat{\beta}(t)\right) = E\left[\left(x^T\hat{\beta}(t) - x^T\beta(t)\right)^2\right] \tag{2.22}$$

Similarly, a useful measure of the global fitness of $\hat{\beta}(t)$ is the average mean squared error (AMSE)

$$\text{AMSE}_\lambda\left(\hat{\beta}\right) = \frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{n_i} E\left[\left(X_{ij}^T\hat{\beta}(t_{ij}) - X_{ij}^T\beta(t_{ij})\right)^2\right]. \tag{2.23}$$

Another quantity which measures the global risk of $\hat{\beta}$ in a similar manner as the AMSE is the average predictive squared error

$$\text{APSE}_\lambda\left(\hat{\beta}\right) = \frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{n_i} E\left[\left(Y_{ij}^* - X_{ij}^T\hat{\beta}(t_{ij})\right)^2\right] \tag{2.24}$$

where Y_{ij}^* is a new observation at (X_{ij}, t_{ij}), i.e. $Y_{ij}^* = X_{ij}^T\beta(t_{ij}) + \epsilon_i^*(t_{ij})$ and $\epsilon_i^*(t)$ is a new mean zero stochastic process which has the same distribution as $\epsilon_i(t)$.

For the data-driven smoothing parameters of this paper, we concentrate on the global fitness of all the components of $\hat{\beta}$. Specifically, we consider the averaged predictive squared error of $\hat{\beta}$. Let $\hat{\beta}^{(-i)}(t)$ be an estimate of $\beta(t)$ based on any one of the linear smoothing methods described in Sections 2.2 through 2.4 by leaving out all the observed measurements of the ith subject. The cross-validation APSE criterion is defined as

$$\text{CV}(\lambda) = \frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\left[Y_{ij} - X_{ij}^T\hat{\beta}^{(-i)}(t_{ij})\right]^2. \tag{2.25}$$

Then our cross-validation smoothing parameter is defined to be λ_{cv} which minimizes CV(λ).

Remark 2.4. For the kernel estimates given in (2.6), minimizing CV(λ) should return one single smoothing parameter, the cross-validation bandwidth h_{cv}. For smoothing splines of Section 2.3, the cross-validation smoothing parameters consist $\lambda_{0,cv}, \ldots, \lambda_{k,cv}$. Similarly, the cross-validation smoothing parameters of LWPEs are window size and possibly the degree of the polynominal used. Intuitively, the extra number of smoothing parameters allowed in smoothing splines can be used to accompany possibly different smoothness of the nonparametric components $\beta_0(t), \ldots, \beta_k(t)$. In particular, if $\beta_0(t), \ldots, \beta_k(t)$ satisfy different smoothness conditions, it may not be possible to obtain appropriate fits to all these components using kernel estimates or LWPEs which only one and possibly two smoothing parameters, respectively. If a single smoothing parameter is desired for the smoothing spline procedure, the covariates should all be standardized, since the λ_j are not dimensionless parameters. ♠

Remark 2.5. In practice, CV(λ) can only be minimized within a preselected compact set of the parameter space and there may be more than one local minima. Thus it is often useful to first try several λ subjectively, and then determine a workable range of λ by examining the fits and the CV(λ) values in this range. In case there is more than one local minima, it is also helpful to re-examine the fit of $\hat{\beta}$ with each λ which gives the corresponding local minimum, instead of using λ which gives the smallest local minima. ♠

Remark 2.6. The effect of leaving out one single subject in the computation of CV(λ) can be explicitly calculated as a perturbation from the full solution. Thus, an alternative expression of CV(λ) maybe useful to speed up the computation. To see how this works for kernel estimates, notice first that, by (2.5),

$$A(t)\hat{\beta}(t) = \sum_{i=1}^{n} J_i(t)Y_i \qquad \text{and} \qquad A^{(-i)}(t)\hat{\beta}^{(-i)}(t) = \sum_{j=1}^{n} J_j(t)Y_j - J_i(t)Y_i$$

where $J_i(t) = X_i^T K_i(t)$, $A(t) = \sum_{i=1}^{n} X_i^T K_i(t) X_i$ and $A^{(-i)}(t) = \sum_{j \neq i} X_j^T K_j(t) X_j$. By the matrix updating formula, i.e. equation (A.2.1) of Cook and Weisberg (1982), we have

$$\begin{aligned} \left(A^{(-i)}(t)\right)^{-1} &= \left(A(t) - X_i^T J_i^T(t)\right)^{-1} \\ &= A^{-1}(t) + A^{-1}(t) X_i^T \left[I - J_i^T(t) A^{-1}(t) X_i^T\right]^{-1} J_i^T(t) A^{-1}(t) \end{aligned} \tag{2.26}$$

where $J_i^T(t) = K_i(t)X_i$ and I is the $n_i \times n_i$ identity matrix. Then (2.25) implies that, for kernel estimates,

$$\mathrm{CV}(h) = \frac{1}{N} \sum_{i=1}^{n} \sum_{j=1}^{n_i} \left[Y_{ij} - X_{ij}^T \left(A^{(-i)}(t_{ij})\right)^{-1} \left(\sum_{i'=1}^{n} J_{i'}(t_{ij}) Y_{i'} - J_i(t_{ij}) Y_i \right) \right]^2$$

where $\left(A^{(-i)}(t_{ij})\right)^{-1}$ can be computed using the right hand side of (2.26). Similar expressions can also be derived for smoothing splines and LWPEs. ♠

3 An Application to Growth of Children

The data considered here involve infants' genders and HIV infection status (HIV positive or negative) measured one year after birth, the third trimester maternal vitamin A levels during pregnancy and repeatedly measured weights of 328 infants from an African AIDS cohort study at the Johns Hopkins University. The covariates in this example are not time varying, although we allow their coefficients to be; further discussion of this structure is contained in Section 5.1. All infants were born from HIV infected mothers in central Africa and survived beyond one year of age. The follow-up study lasted two years and infants' weights were recorded during every scheduled monthly visit. Due to various reasons, a number of the scheduled visits were missed by some infants which resulted in unequal numbers of repeated weight measurements per infant. The main objective is to evaluate the time-varying effects of two binary covariates, child's gender and HIV status, and one continuous covariate, the third trimester maternal vitamin A level, on child's weight. Previous studies have shown that vitamin A can improve immune function and resistance to disease [cf. Semba (1994)]. Biologically, a significant association between maternal vitamin A levels and infant growth may suggest the benefit of vitamin A supplementation in the mother's and infant's diet.

This data set was initially analyzed by Semba et al. (1996) where vitamin A was treated as a binary covariate (deficiency and nondeficiency), and the growth prediction curves were obtained using parametric regression models and generalized estimation equations [cf. Diggle, Liang and Zeger (1994)]. Here we use the actual measurements and fit the data to (1.2) with $X_{i10} = \cdots = X_{in_i0} = 1$,

$$X_{i11} = \cdots = X_{in_i1} = \begin{cases} 1 & \text{if the } i\text{th infant is HIV positive,} \\ 0 & \text{if the } i\text{th infant is HIV negative,} \end{cases}$$

$$X_{i12} = \cdots = X_{in_i2} = \text{ the } i\text{th infant's maternal vitamin A level,}$$

$$X_{i13} = \cdots = X_{in_i3} = \begin{cases} 1 & \text{if the } i\text{th infant is male,} \\ 0 & \text{if the } i\text{th infant is female,} \end{cases}$$

$$Y_{ij} = \text{ weight in kilograms of the } i\text{th infant at time } t_{ij} \text{ after birth.}$$

For brevity, we only present the smoothing results of kernel and spline methods. The smoothing results of LWPEs are very similar to those given by kernels and smoothing splines.

Figure 1 shows the estimated values of $\beta_l(t)$, $l = 0, \ldots, 3$, together with their ± 2 pointwise bootstrap standard error bands, the weight prediction curves with the maternal vitamin A level taken to be $29.5\mu g/ml$ (the mean vitamin A level in the sample), and the plot of residuals vs. time points. A Gaussian kernel with $h = 1.2$ was used. The bootstrap standard errors were computed at 100 selected time points using 200 bootstrap replications by resampling from the subjects, i.e. randomly resample the entire repeated measurements of the subjects with replacement. From the figure it is seen that the magnitudes of the coefficients of all three factors initially increase with time and then level off. The initial increase with time probably reflects the cumulative effects of additional diseases early in life due to HIV infection and/or low vitamin A levels. The leveling off of the difference may be due to the establishment of the infants immunity function at one year of age and frailty effects from the sickest and lowest weight babies dyeing. It appears that the weight prediction at birth (a boundary point) has small amount of bias, but the predictions at all interior time points are quite reasonable. All the estimates, $\hat{\beta}_l(t)$, $l = 0, \ldots, 3$, appear to be slightly under-smoothed. The residual plot reveals a generally nice pattern despite a slight increase of variation when $t \geq 2$.

Figure 2 shows the similar graphs as in Figure 1 except that $h = 1.2$ is replaced by the cross-validation bandwidth $h_{cv} = 0.5$. Although the cross-validation results give a slightly better weight prediction at birth, all other estimates and predictions appear to be substantially undersmoothed. Thus, because of the small bandwidth, the underlying patterns of the coefficients $\beta_0(t)$, $\beta_1(t)$, $\beta_2(t)$ and $\beta_3(t)$ can not be easily visualized. In view of Remarks 2.5 and 2.6, a possible cause of this severe undersmoothness is that the cross-validation criterion concentrated on minimizing the prediction errors.

As a comparison to the kernel estimates of Figures 1 and 2, Figure 3 shows the results of natural cubic splines when the smoothing parameters, λ_l, $l = 0, 1, 2, 3$, are chosen to be 0.7, 0.7, 20 and 0.7, respectively. All the estimates of $\beta_l(t)$ and the prediction curves give generally the same patters as those obtained in Figure 1. Despite a large smoothing parameter used for $\beta_2(t)$, the estimated curve $\hat{\beta}_2(t)$ still appears to be a little undersmoothed. More substantial undersmoothing would appear if λ_2 were also chosen to be 0.7.

Figure 4 presents the similar graphs as in Figure 3 except that the cross-validation smoothing parameters 0.125, 0.2, 2.5 and 0.2 are used for λ_0, λ_1, λ_2 and λ_3, respectively. Although λ_0, λ_1 and λ_3 are smaller than those used in Figure 3, the general shapes of $\hat{\beta}_0(t)$, $\hat{\beta}_1(t)$, and $\hat{\beta}_3(t)$ are basically the same as the corresponding estimates given in Figure 3. Again, the cross-validation estimate of $\beta_2(t)$ seems to be slightly undersmoothed. It is interesting to note here that the cross-validation procedure for smoothing splines gives different λ_l values for $l = 0, \ldots, 3$.

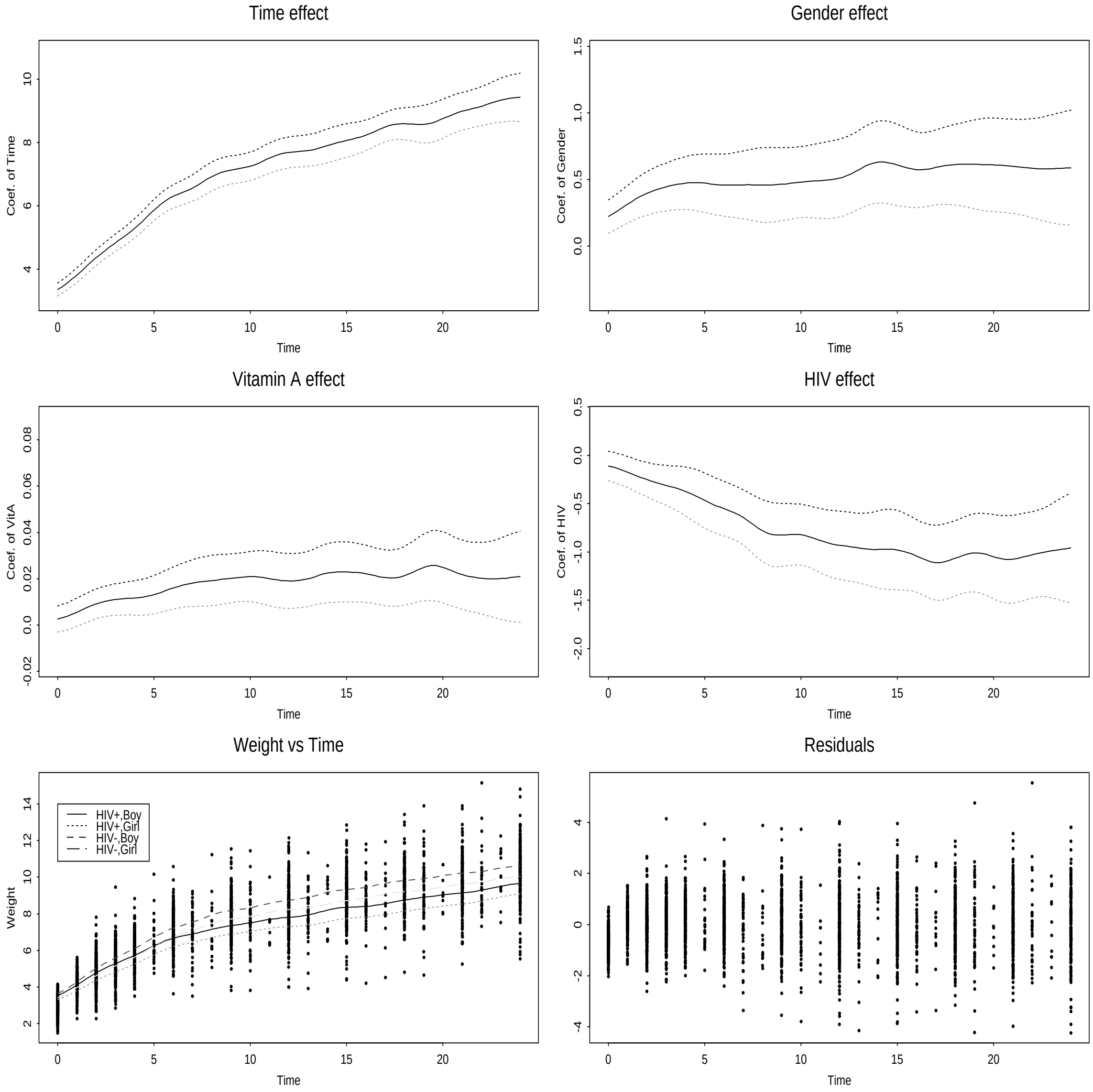

Figure 1: Estimates, predictions and residuals using kernel method with the standard Gaussian kernel and $h = 1.2$ as the bandwidth. The dashed curves in the top four graphs of the estimates represent the ± 2 bootstrap standard error bands. *Time effect*: $\hat{\beta}_0(t)$ vs. time. *HIV effect*: $\hat{\beta}_1(t)$ vs. time. *VitA effect*: the estimated effect of vitamin A $\hat{\beta}_2(t)$ vs. time. *Gender effect*: $\hat{\beta}_3(t)$ vs. time. *Weight vs. Time*: The weight prediction curve when the vitamin A level is $29.5\mu g/ml$. "·" represents actual data. *Residuals*: Plot of residuals vs. time.

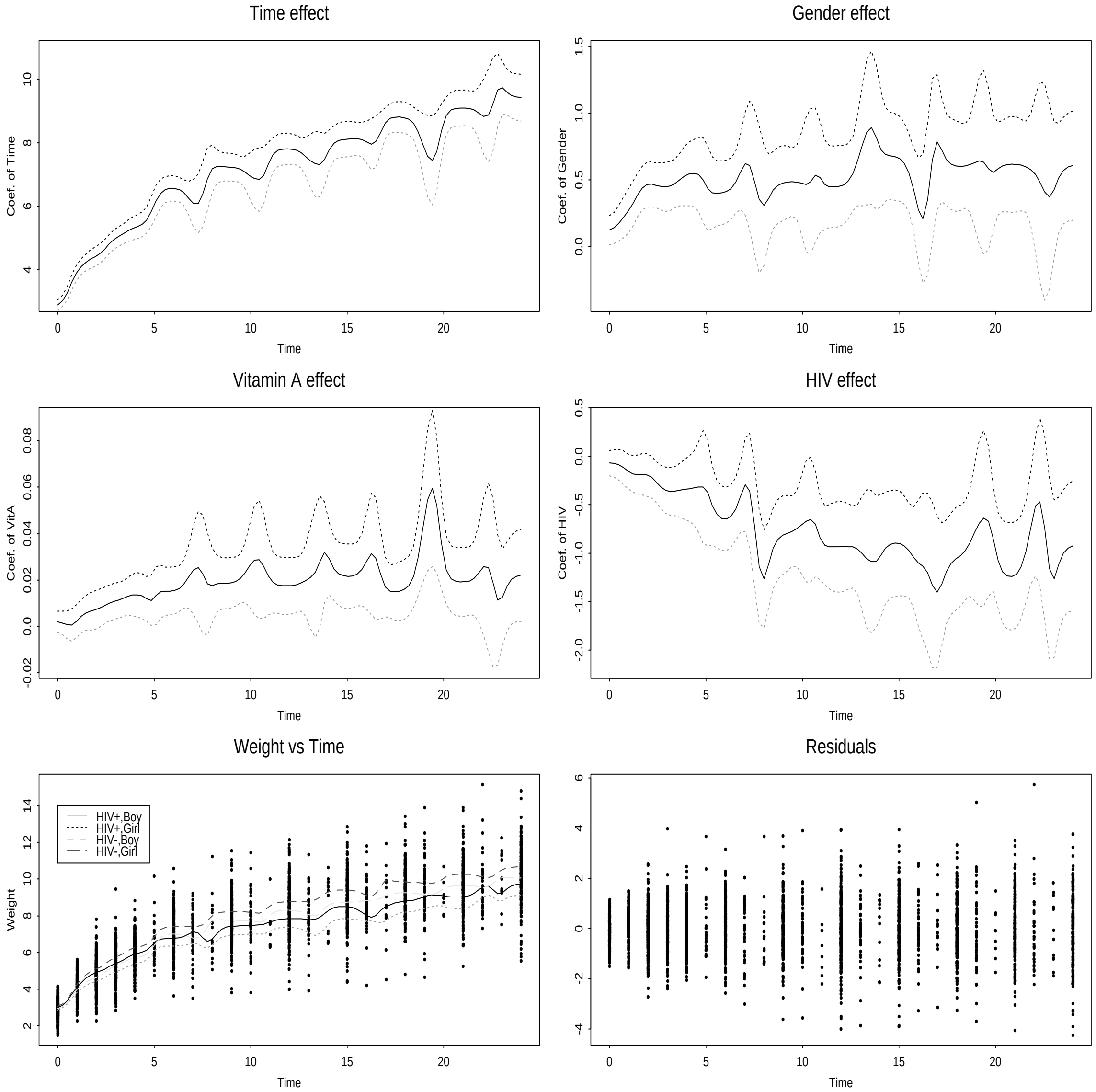

Figure 2: Estimates, predictions and residuals using kernel method with the standard Gaussian kernel and the cross-validation bandwidth $h_{cv} = 0.5$. The dashed curves in the top four graphs of the estimates represent the ± 2 bootstrap standard error bands. *Time effect*: $\hat{\beta}_0(t)$ vs. time. *HIV effect*: $\hat{\beta}_1(t)$ vs. time. *VitA effect*: the estimated effect of vitamin A $\hat{\beta}_2(t)$ vs. time. *Gender effect*: $\hat{\beta}_3(t)$ vs. time. *Weight vs. Time*: The weight prediction curve when the vitamin A level is $29.5\mu g/ml$. "·" represents actual data. *Residuals*: Plot of residuals vs. time.

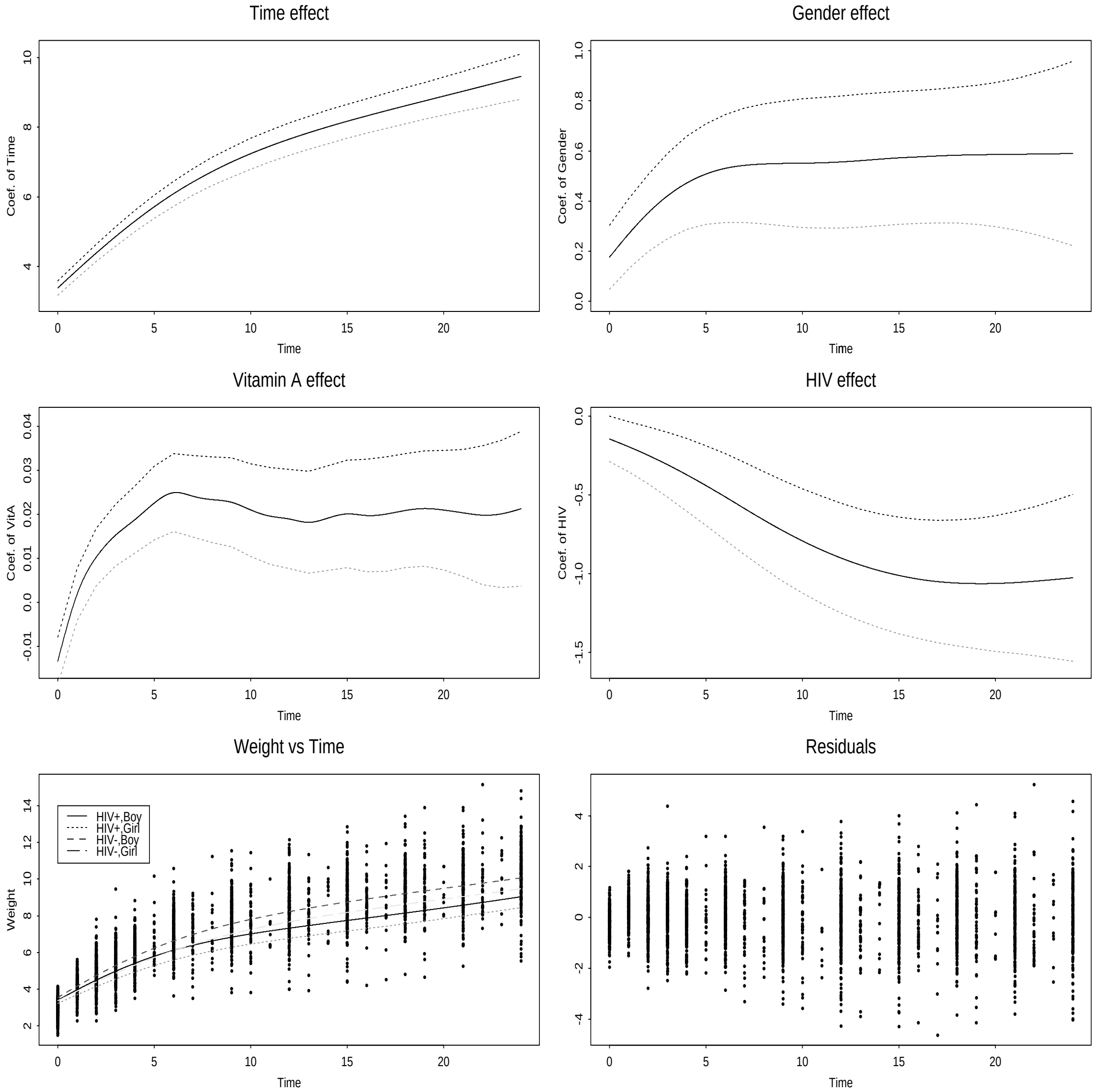

Figure 3: Estimates, predictions and residuals using natural cubic splines with 0.7, 0.7, 20 and 0.7 as the corresponding smoothing parameters λ_l, $l = 0, \ldots, 3$, respectively. The dashed curves in the top four graphs of the estimates represent the ± 2 bootstrap standard error bands. *Time effect*: $\hat{\beta}_0(t)$ vs. time. *HIV effect*: $\hat{\beta}_1(t)$ vs. time. *VitA effect*: the estimated effect of vitamin A $\hat{\beta}_2(t)$ vs. time. *Gender effect*: $\hat{\beta}_3(t)$ vs. time. *Weight vs. Time*: The weight prediction curve when the vitamin A level is $29.5\mu g/ml$. "·" represents actual data. *Residuals*: Plot of residuals vs. time.

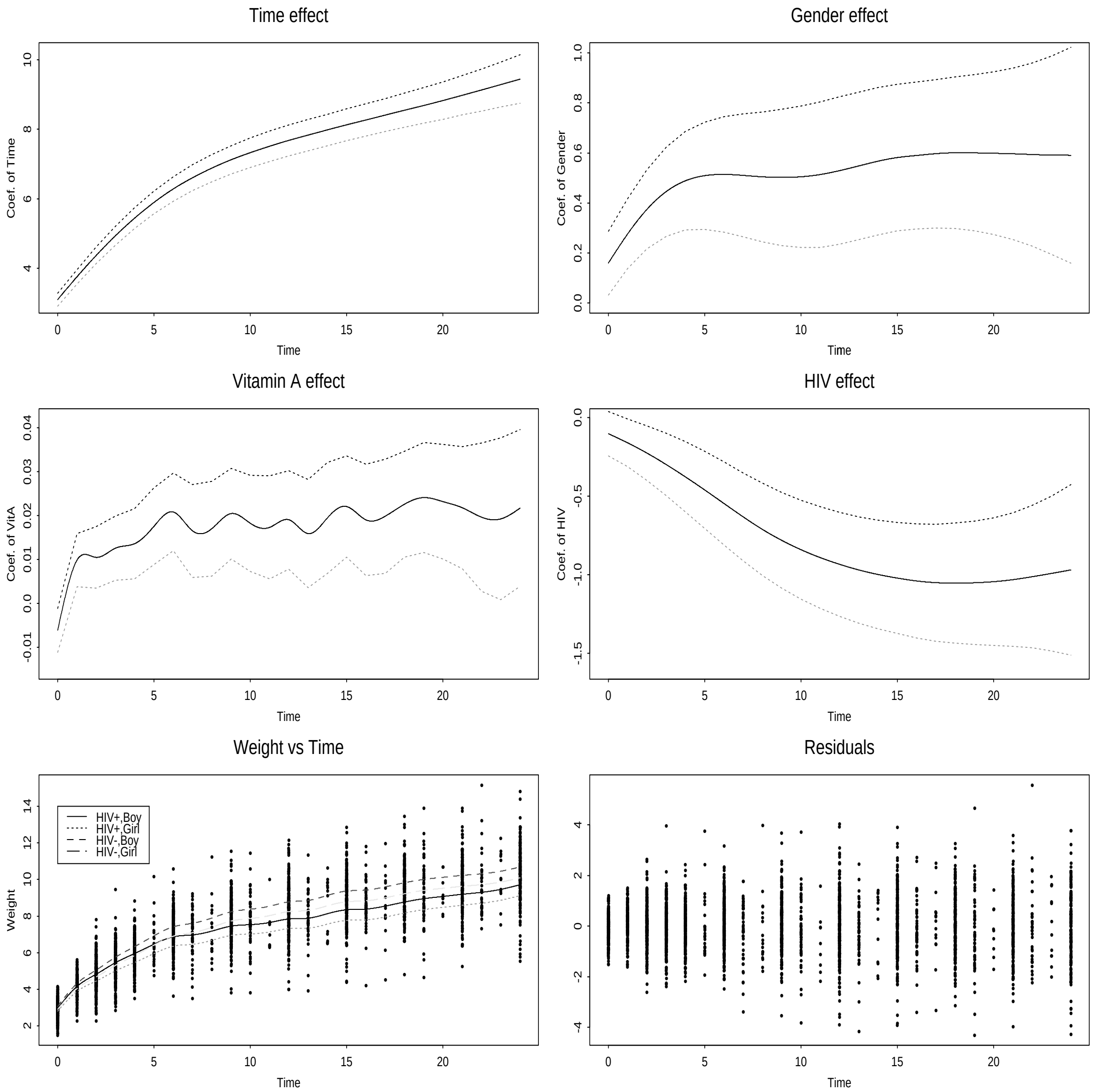

Figure 4: Estimates, predictions and residuals using natural cubic splines with λ_l, $l = 0, \ldots, 3$ chosen to be the cross-validation smoothing parameters 0.125, 0.2, 2.5 and 0.2, respectively. The dashed curves in the top four graphs of the estimates represent the ± 2 bootstrap standard error bands. *Time effect*: $\hat{\beta}_0(t)$ vs. time. *HIV effect*: $\hat{\beta}_1(t)$ vs. time. *VitA effect*: the estimated effect of vitamin A $\hat{\beta}_2(t)$ vs. time. *Gender effect*: $\hat{\beta}_3(t)$ vs. time. *Weight vs. Time*: The weight prediction curve when the vitamin A level is $29.5\mu g/ml$. "·" represents actual data. *Residuals*: Plot of residuals vs. time.

4 Asymptotic Risks of Kernel Estimates

A natural approach to evaluate the adequacy of an estimator is through the asymptotic behaviors of its risks. For mathematical simplicity, we only consider in this section the asymptotic risk representations under the mean squared errors for the kernel estimates as defined in (2.6).We believe that the asymptotic risks of smoothing splines and locally weighted polynomials (LWPEs) may be analogously investigated, but we have not done so.

4.1 Assumptions and Mean Squared Errors

The estimation methods of Section 2 can accommodate both fixed and random designs. For technical convenience, we assume that the design points t_{ij}, $j = 1, \ldots, n_i$ and $i = 1, \ldots, n$, are chosen independently according to some design distribution F_T and design density f_T. (An examination of the proofs will reveal that the assumption that the design points within subjects are chosen in such a way can be substantially relaxed.) In addition, the following technical conditions are assumed throughout this section:

A1. *X_{ij} and $\epsilon_i(\cdot)$ are independent for all $i = 1, \ldots, n$ and $j = 1, \ldots, n_i$. In addition,*

$$E\left[\epsilon^2(t)\right] < \infty \qquad \text{and} \qquad \lim_{\Delta \to 0} E\left[\epsilon(t + \Delta)\, \epsilon(t)\right] < \infty.$$

A2. *Let $E^{(lr)}_{XX^T}(s)$ be the lrth component of the matrix $E\left(X(s)X^T(s)\right)$. Each $E^{(lr)}_{XX^T}(s)$ with $l, r = 0, \ldots, k$ is Lipschitz continuous in the sense that there are positive constants c_0 and α_0 such that, for all $s_1, s_2 \in R$,*

$$\left|E^{(lr)}_{XX^T}(s_1) - E^{(lr)}_{XX^T}(s_2)\right| \leq c_0 |s_1 - s_2|^{\alpha_0}.$$

Furthermore, $E\left(X^4_{ijl}\right) < \infty$ for all $i = 1, \ldots, n$, $j = 1, \ldots, n_i$ and $l = 0, \ldots, k$.

A3. *$\beta_l(\cdot)$, $l = 0, \ldots, k$, are Lipschitz continuous in the sense that there are positive constants c_1 and α_1 such that, for all $t_1, t_2 \in R$,*

$$|\beta_l(t_1) - \beta_l(t_2)| \leq c_1 |t_1 - t_2|^{\alpha_1}.$$

A4. *The design density $f_T(\cdot)$ is bounded away from zero at t, i.e. $f_T(t) \geq b$ for some $b > 0$, and is Lipschitz continuous, i.e., there are positive constants c_2 and α_2 such that, for all $t_1, t_2 \in R$,*

$$|f_T(t_1) - f_T(t_2)| \leq c_2 |t_1 - t_2|^{\alpha_2}.$$

A5. The kernel function $K(\cdot)$ is bounded on R, and satisfies

$$\int u^{\alpha} K(u) < \infty, \qquad \int K(u)\, du = 1 \qquad \textit{and} \qquad \int K^2(u)\, du < \infty$$

where $\alpha = \max(\alpha_0, \alpha_1, \alpha_2)$.

A6. The bandwidth h depends on n and satisfies $\lim_{n\to\infty} h = 0$ *and* $\lim_{n\to\infty} nh = \infty$.

These assumptions are comparable with the regularity conditions commonly used in nonparametric regression under independent cross-sectional data, e.g. Härdle (1990), and are general enough to be satisfied in many interesting practical situations. Theoretically, these assumptions could be further modified or even weakened in various ways so that more desirable asymptotic properties of the kernel estimate $\hat\beta(t)$ may be obtained. Some of these modifications and special cases will be further discussed later in this Section.

The risk of $\hat\beta(\cdot)$ as an estimate of $\beta(\cdot)$ depends on the choice of loss functions. As mentioned in Section 2.5, because of its mathematical tractability a popular choice is squared loss. Suppose that only the local risks of $\hat\beta(\cdot)$ at time t are considered. Then the mean squared errors as defined in (2.20) and (2.22) are reasonable measures of the risk of the lth component $\hat\beta_l(t)$ and the predictive risk of $\hat\beta(t)$, respectively. In general, (2.20) and (2.22) are special cases of the following risk function

$$\text{(4.1)} \qquad \text{MSE}_{\mathbf{w}}\left(\hat\beta(t)\right) = E\left[\left(\hat\beta(t) - \beta(t)\right)^T \mathbf{w} \left(\hat\beta(t) - \beta(t)\right)\right]$$

where $\mathbf{w} = diag(w_0, \ldots, w_k)$ with non-negative diagonal elements w_l, $l = 1, \ldots, k$. Here, (2.20) corresponds to the case of $w_l = 1$ and $w_r = 0$ for all $r \neq l$, and (2.21) corresponds to the case of $w_l = x_l^2$, $l = 0, \ldots, k$.

Unfortunately, a minor technical inconvenience for the kernel estimate $\hat\beta(t)$ is that its mean squared error as defined in (4.1) may not exist in general [cf. Rosenblatt (1969) and Härdle and Marron (1983)]. Thus, a slight modification of (4.1) has to be considered. By (2.3) and (2.6), we have

$$\begin{aligned} \hat\beta(t) - \beta(t) \quad &= \quad \left(\sum_{i=1}^{n} X_i^T K_i(t) X_i\right)^{-1} \left(\sum_{i=1}^{n} X_i^T K_i(t) Y_i\right) \\ &\quad - \left(E\left[X(t) X^T(t)\right]\right)^{-1} \left(E\left[X(t) Y(t)\right]\right). \end{aligned}$$

Rearranging the terms on both sides, the above equation implies

$$\begin{aligned} \text{(4.2)} \quad (1 + \Delta(t))\left(\hat\beta(t) - \beta(t)\right) \quad &= \quad f_T^{-1}(t)\left(E_{XX^T}(t)\right)^{-1} \left[\left(\frac{1}{Nh}\sum_{i=1}^{n} X_i^T K_i(t) Y_i\right)\right. \\ &\quad \left. - \left(\frac{1}{Nh}\sum_{i=1}^{n} X_i^T K_i(t) X_i\right)\beta(t)\right] \end{aligned}$$

where

$$\Delta(t) = f_T^{-1}(t)\,(E_{XX^T}(t))^{-1}\left[\frac{1}{Nh}\sum_{i=1}^{n} X_i^T K_i(t) X_i - E_{XX^T}(t)\,\hat{f}_T(t)\right] + f_T^{-1}(t)\left(\hat{f}_T(t) - f_T(t)\right),$$

$$\hat{f}_T(t) = \frac{1}{Nh}\sum_{i=1}^{n}\sum_{j=1}^{n_i} K\left(\frac{t-t_{ij}}{h}\right)$$

and $E_{XX^T}(t)$ denotes the $(k+1)\times(k+1)$ matrix $E\left[X(t)X^T(t)\right]$. Notice that $\hat{f}_T(t)$ is a kernel estimate of the underlying marginal density $f_T(t)$. It can be easily deduced from standard results of kernel density estimates [e.g. Härdle (1990)] that $\hat{f}_T(t) - f(t) = o_p(1)$ as $n \to \infty$ and $h \to 0$. Lemma 1 of Section 6 shows that, by taking $a_{i,j,l_1,l_2}(t)$ of Lemma 1 to be $X_{ijl_1}X_{ijl_2} - E(X_{l_1}(t)X_{l_2}(t))$, all the (l,r)th elements of the $(k+1)\times(k+1)$ matrix

$$\frac{1}{Nh}\sum_{i=1}^{n} X_i^T K_i(t) X_i - E_{XX^T}(t)\,\hat{f}_T(t)$$

converges to zero in probability as $n \to \infty$ and $h \to 0$. Then, a direct consequence of (4.2) is

$$(4.3)\qquad (1 + o_p(1))\left(\hat{\beta}(t) - \beta(t)\right) = f_T^{-1}(t)\,(E_{XX^T}(t))^{-1}\,\hat{R}(t)$$

where

$$\hat{R}(t) = \left(\frac{1}{Nh}\sum_{i=1}^{n} X_i^T K_i(t) Y_i\right) - \left(\frac{1}{Nh}\sum_{i=1}^{n} X_i^T K_i(t) X_i\right)\beta(t).$$

Equation (4.3) implies that the error term $\hat{\beta}(t) - \beta(t)$ can be approximated in probability by $f_T^{-1}(t)\,(E_{XX^T}(t))^{-1}\,\hat{R}(t)$.

Thus, to avoid the technical inconvenience that might arise due to nonexistence of the mean squared errors, the asymptotic risk of $\hat{\beta}(t)$ is described through the modified mean squared error

$$(4.4)\qquad \mathrm{MSE}^*_{\mathbf{w}}\left(\hat{\beta}(t)\right) = E\left[\hat{R}^T(t) M(t) \hat{R}(t)\right] = \sum_{l=0}^{k}\sum_{r=0}^{k}\left\{M_{lr}(t)\,E\left[\hat{R}_l(t)\hat{R}_r(t)\right]\right\}$$

where $\hat{R}_l(t)$ is the lth element of the $k+1$ column vector $\hat{R}(t)$ and $M_{lr}(t)$ is the (l,r)th element of the $(k+1)\times(k+1)$ matrix

$$M(t) = f_T^{-2}(t)\left((E_{XX^T}(t))^{-1}\right)^T \mathbf{w}\,(E_{XX^T}(t))^{-1}.$$

Similarly, the bias of $\hat{\beta}(t)$ can be measured by

$$(4.5)\qquad \mathrm{B}^*\left(\hat{\beta}(t)\right) = f_T^{-1}(t)\,(E_{XX^T}(t))^{-1}\,E\hat{R}(t).$$

4.2 Asymptotic Risk Representations

For the simple case that only the risk of one component, say $\hat{\beta}_l(t)$, is considered, the mean squared error $\mathrm{MSE}^*_{\mathbf{w}}(\hat{\beta}(t))$ only depends on the second moment of $\hat{R}_l(t)$ which can be computed by evaluating the expectations, variances and covariances of the corresponding terms of $\hat{R}_l(t)$. But, more generally, if two or more elements of the diagonal matrix $\mathbf{w}$ are strictly positive, $\mathrm{MSE}^*_{\mathbf{w}}(\hat{\beta}(t))$ involves more complicated terms, such as expectations of the cross products of different elements of $\hat{R}(t)$. Thus, some notation is needed. Let

$$
\begin{aligned}
\sigma^2(t) &= E\left[\epsilon_i^2(t)\right], \qquad i = 1, \ldots, n, \\
\rho_\epsilon(t) &= \lim_{\Delta \to 0} E\left[\epsilon_i(t+\Delta)\,\epsilon_i(t)\right], \qquad i = 1, \ldots, n, \\
\rho_{lr}(s) &= E\left[X_{ijl} X_{ijr} | t_{ij} = s\right], \\
B_l(t) &= \sum_{r=0}^{k} \int \left(\beta_r(t-hu) - \beta_r(t)\right) \rho_{lr}(t-hu) f_T(t-hu) K(u)\, du, \\
\mathrm{I}_N(t) &= \sum_{l_1=0}^{k} \sum_{l_2=0}^{k} \left[M_{l_1 l_2}(t) B_{l_1}(t) B_{l_2}(t)\right], \\
\mathrm{II}_N(t) &= N^{-2} \rho_\epsilon(t) N^{-2} \left(\sum_{i=1}^{n} n_i^2 - N\right) \left[\sum_{l_1=0}^{k} \sum_{l_2=0}^{k} M_{l_1 l_2}(t)\, \rho_{l_1 l_2}(t)\right], \\
\mathrm{III}_N(t) &= (Nh)^{-1}\, \sigma^2(t) \left[\sum_{l_1=0}^{k} \sum_{l_2=0}^{k} M_{l_1 l_2}(t) \rho_{l_1 l_2}(t)\right].
\end{aligned}
$$

We now summarize the main results of this section in Theorem 1 and three corollaries.

THEOREM 1. *Suppose that assumptions A1 through A6 are satisfied and t is an interior point of the support of f_T. When the number of subjects n is sufficiently large, then the bias and the mean squared error are given by*

$$
\mathrm{B}^*\left(\hat{\beta}(t)\right) = f_T^{-1}(t) \left[E_{XX^T}(t)\right]^{-1} E\left(\hat{R}(t)\right) \tag{4.6}
$$

where $E\hat{R}(t) = \left(B_0(t), \ldots, B_k(t)\right)^T$, *and*

$$
\begin{aligned}
\mathrm{MSE}^*_{\mathbf{w}}\left(\hat{\beta}(t)\right) &= \frac{\sum_{i=1}^{n}\left[n_i\left(\sum_{i \neq i'} n_{i'}\right)\right]}{N^2} \mathrm{I}_N(t) + f_T^2(t) \mathrm{II}_N(t) \\
&\quad + f_T(t) \left(\int K^2(u)\, du\right) \mathrm{III}_N(t) \\
&\quad + o\left(N^{-2}\left(\sum_{i=1}^{n} n_i^2 - N\right)\right) + o\left(N^{-1} h^{-1}\right).
\end{aligned} \tag{4.7}
$$

Furthermore, $lim_{n\to\infty} \mathrm{MSE}^*_{\mathbf{w}}\left(\hat{\beta}(t)\right) = 0$ *if and only if* $lim_{n\to\infty} \max_{1 \le i \le n}(n_i/N) = 0$. □

It can be seen from Theorem 1 that $\mathrm{I}_N(t)$ only involves the biases of $\hat\beta_l(t)$ while the covariances and variances of $\hat\beta_l(t)$ are involved in $\mathrm{II}_N(t)$ and $\mathrm{III}_N(t)$, respectively. The convergence rates of $\mathrm{I}_N(t)$ and the bias $\mathrm{B}^*\left(\hat\beta(t)\right)$ depend on the smoothness of the underlying functions β_r, ρ_{lr} and f_T. However, $\mathrm{II}_N(t)$ and $\mathrm{III}_N(t)$ converge to zero with rates $N^{-2}\left(\sum_{i=1}^n n_i^2 - N\right)$ and $N^{-1}h^{-1}$, respectively. It is interesting to see from (4.7) that the intra-correlations of longitudinal data are only included in the term $\mathrm{II}_N(t)$ in the asymptotic expression of $\mathrm{MSE}^*_{\mathbf{w}}\left(\hat\beta(t)\right)$. Without this extra term, the asymptotic behaviors of $\hat\beta(t)$ would be the same as with independent cross-sectional samples. Theorem 1 further indicates that, in order to ensure the consistency of kernel estimates, no individual or a small finite set of individuals should dominate in terms of proportions of measurements.

For some subfamilies of the smooth functions satisfying A2 through A4, the following interesting results are direct consequences of Theorem 1.

Corollary 1. *Suppose that the assumptions of Theorem 1 are satisfied, β_r, ρ_{lr} and f_T, $l, r = 0, \ldots, k$ are $d \ge 2$ times continuously differentiable and $K(\cdot)$ is a dth-order kernel as defined in Remark 2.1. Let $A^{(a)}(t)$ be the ath derivative of any function $A(t)$ and*

$$b_l(t) = \sum_{r=0}^{k}\sum_{a=0}^{d-1}\sum_{b=0}^{a}\left\{\frac{\beta_r^{(d-a)}(t)\rho_{lr}^{(a-b)}(t)f_T^{(b)}(t)}{(d-a)!\,(a-b)!\,b!}(-1)^d\left(\int u^d K(u)\,du\right)\right\}. \tag{4.8}$$

When n is sufficiently large, the bias and the mean squared error of $\hat\beta(t)$ are given by

$$\mathrm{B}^*\left(\hat\beta(t)\right) = f_T^{-1}(t)\left[E_{XX^T}(t)\right]^{-1}h^d\left(b_0(t),\ldots,b_k(t)\right)^T + o\left(h^d\right) \tag{4.9}$$

and

$$\begin{aligned}\mathrm{MSE}^*_{\mathbf{w}}\left(\hat\beta(t)\right) &= \left(1-\frac{\sum_{i=1}^n n_i^2}{N^2}\right)h^{2d}\left(\sum_{l_1=0}^{k}\sum_{l_2=0}^{k}\left[M_{l_1 l_2}(t)b_{l_1}(t)b_{l_2}(t)\right]\right)\\ &\quad + f_T^2(t)\,\mathrm{II}_N(t) + f_T(t)\left(\int K^2(u)\,du\right)\mathrm{III}_N(t)\\ &\quad + o\left(N^{-2}\left(\sum_{i=1}^n n_i^2 - N\right)\right) + o\left(N^{-1}h^{-1}\right) + o\left(h^{2d}\right).\ \square\end{aligned} \tag{4.10}$$

Corollary 2. *Suppose that, in addition to the assumptions of Corollary 1,*

$$\lim_{n\to\infty}\frac{\sum_{i=1}^n n_i^2}{N} \le \lambda \qquad \textit{for some constant } \lambda > 0. \tag{4.11}$$

The mean squared error of $\hat\beta(t)$ is then given by

$$\mathrm{MSE}^*_{\mathbf{w}}\left(\hat\beta(t)\right) \tag{4.12}$$

$$
\begin{aligned}
&= h^{2d}\left(\sum_{l_1=0}^{k}\sum_{l_2=0}^{k}\left[M_{l_1 l_2}(t) b_{l_1}(t) b_{l_2}(t)\right]\right)\\
&\quad + (Nh)^{-1} f_T(t)\left(\int K^2(u)\,du\right)\sigma^2(t)\left[\sum_{l_1=0}^{k}\sum_{l_2=0}^{k} M_{l_1 l_2}(t)\rho_{l_1 l_2}(t)\right]\\
&\quad + o\left(h^{2d}\right) + o\left(N^{-1}h^{-1}\right). \ \square
\end{aligned}
$$

COROLLARY 3. *Suppose that the assumptions of Corollary 2 are satisfied. The optimal bandwidth which minimizes the dominating terms of* $\mathrm{MSE}^*_{\mathbf{w}}\left(\hat\beta(t)\right)$ *is*

$$
h_{opt} = N^{-\frac{1}{2d+1}}\left[\frac{f_T(t)\left(\int K^2(u)\,du\right)\sigma^2(t)\left(\sum_{l_1=0}^{k}\sum_{l_2=0}^{k} M_{l_1 l_2}(t)\rho_{l_1 l_2}(t)\right)}{2d\sum_{l_1=0}^{k}\sum_{l_2=0}^{k} M_{l_1 l_2}(t) b_{l_1}(t) b_{l_2}(t)}\right]^{\frac{1}{2d+1}}, \tag{4.13}
$$

and, by substituting h_{opt} *into (4.11), the optimal mean squared error is given by*

$$
\begin{aligned}
\mathrm{MSE}^*_{\mathbf{w}}\left(\hat\beta_{h_{opt}}(t)\right) &= N^{\frac{-2d}{2d+1}}\left[(2d)^{\frac{-2d}{2d+1}} + (2d)^{\frac{1}{2d+1}}\right]\\
&\quad\times\left[f_T(t)\left(\int K^2(u)\,du\right)\sigma^2(t)\left(\sum_{l_1=0}^{k}\sum_{l_2=0}^{k} M_{l_1 l_2}(t)\rho_{l_1 l_2}(t)\right)\right]^{\frac{2d}{2d+1}}\\
&\quad\times\left[\sum_{l_1=0}^{k}\sum_{l_2=0}^{k} M_{l_1 l_2}(t) b_{l_1}(t) b_{l_2}(t)\right]^{\frac{1}{2d+1}}\\
&\quad + o\left(N^{\frac{-2d}{2d+1}}\right). \ \square
\end{aligned} \tag{4.14}
$$

One implication of the above results is that the asymptotic behaviors of $\hat\beta(t)$ also depend on how fast n and n_i, $i = 1, \ldots, n$, converge to infinity. In general, $\hat\beta(t)$ is not necessarily a consistent estimate of $\beta(t)$ when only N converges to infinity. For example, if $n_i = m$, $i = 1, \ldots, n$, m converges to infinity but n stays bounded, then, since $N^{-2}\left(\sum_{i=1}^{n} n_i^2 - N\right) = n^{-1} - N^{-1}$ is bounded away from zero for sufficiently large N, $\mathrm{MSE}^*_{\mathbf{w}}\left(\hat\beta(t)\right)$ does not converge to zero as N goes to infinity. But in most practical applications, n_i stay bounded for all $i = 1, \ldots, n_i$, and n converges to infinity, then by Corollary 3 $\hat\beta(t)$ is asymptotically equivalent to that with independent cross-sectional samples.

Remark 4.1. Asymptotically the effect of the intra-correlations on $\mathrm{MSE}^*_{\mathbf{w}}\left(\hat\beta(t)\right)$ appears in $\mathrm{II}_N(t)$ through the covariance term $\rho_\epsilon(t)$ of $\epsilon_i(t)$. In general, $\rho_\epsilon(t)$ does not necessary equal to $\sigma^2(t)$ [cf. Zeger and Diggle (1994)]. Notice here that $\mathrm{II}_N(t)$ only depends on the limiting values, $\rho_\epsilon(t)$, of the covariances of $\epsilon_i(t)$ and $\epsilon_i(s)$ as $s \to t$. This is because of the local averaging nature of kernel methods: the estimates tend to ignore the measurements at design points t_{ij} which are outside a shrinking neighborhood of t. Since the bandwidths shrink to zero, any correlation between $\epsilon_i(t)$ and $\epsilon_i(s)$, $t \neq s$, is ignored when n is sufficiently large. This is fortunate, since in practice we may only aware the presence of the intra-correlations but have very little knowledge about the specific correlation structures. By using a kernel estimate or any other equivalent local smoothing method, we essentially choose to ignore the correlation structures. ♠

Remark 4.2. The reason that the ideal optimal bandwidth of (4.13) can be explicitly derived from (4.12) is because the second term at the right hand side of (4.7) converges to zero much faster than $N^{-1}h^{-1}$ when A6 is satisfied and N goes to infinity. In particular, when $d = 2$, Corollary 3 shows that h_{opt} should be of the order $N^{-1/5}$. Then the best possible convergence rate for $\hat{\beta}(t)$ is $N^{-4/5}$ in terms of the mean squared errors. As it is usually the case in nonparametric regression, h_{opt} depends on the unknown functions, such as $f_T(t)$, $\rho_{l_1 l_2}(t)$ and $b_l(t)$, etc., so that a data-driven estimate of h_{opt} is needed in practice. We conjecture that some plug-in procedures analogous to those discussed in Härdle (1990) or the cross-validation of Section 2.5 should be reasonable candidates to obtain good estimates of h_{opt}. But the theoretical justification of this conjecture is beyond the scope of this paper and is omitted. ♠

Remark 4.3. Different smoothing parameters might be needed in practice to accommodate different smoothness among $\beta_0(t), \ldots, \beta_k(t)$. The kernel estimate $\hat{\beta}(t)$ as defined in (2.6) only involves one smoothing parameter, hence, is not adequately equipped for this task. For mathematical simplicity, assumptions A2 through A4 require that all the underlying functions involved in $\beta(t)$ belong to the same smoothness family, which may not be realistic in many situations. Even so, the restriction of using one single smoothing parameter in $\hat{\beta}(t)$ may still produce undesirable smoothing results. On the other hand, the smoothing splines of Section 2.3 involve $k+1$ smoothing parameters for the corresponding $k+1$ nonparametric functions to be estimated. Thus further comparisons of theoretical and practical adequacy between different linear estimates may be revealing. ♠

5 Discussion

In this section, we briefly discuss some extensions of model (1.2) and some open problems. A number of potentially interesting topics merit further investigation.

5.1 Time Independent Covariates

In many situations such as the epidemiological study of Section 3, the covariates X do not depend on t, and only the outcome variable Y is repeatedly measured. Then the expectation $E\left(X(t)X^T(t)\right)$ equals the $(k+1)\times(k+1)$ matrix $E\left(XX^T\right)$. Assuming that $E\left(XX^T\right)$ is invertible, (2.3) reduces to

$$\beta(t) = E\left(XX^T\right)^{-1} E\left(XY(t)\right). \tag{5.1}$$

Denoting the observed covariates of the ith subject by $X_{i0}, \ldots, X_{ik}$ where $X_{il} \in R$, $l = 0, \ldots, k$, an obvious estimate of the $E\left(XX^T\right)$ is the sample average

$$\tilde{E}_{XX^T} = n^{-1}\sum_{i=1}^{n}\left(X_{i0}, \ldots, X_{ik}\right)^T\left(X_{i0}, \ldots, X_{ik}\right).$$

Equation (5.1) suggests that smoothing is then only needed for the estimation of $E\left(XY(t)\right)$. Suppose that one would like to estimate $\beta(t)$ using a kernel estimate. The modified version of (2.6) then becomes

$$\tilde{\beta}(t) = \left(\tilde{E}_{XX^T}\right)^{-1}\left(N^{-1}h^{-1}\sum_{i=1}^{n}\tilde{X}_i^T K_i(t) Y_i\right) \tag{5.2}$$

where Y_i and $K_i(t)$ are as defined in Section 2.2 and

$$\tilde{X}_i = \begin{pmatrix} X_{i0} & X_{i1} & \cdots & X_{ik} \\ \cdots & \cdots & \cdots & \cdots \\ X_{i0} & X_{i1} & \cdots & X_{ik} \end{pmatrix}.$$

The asymptotic properties of $\tilde{\beta}(t)$ then depend on the large sample behavior of

$$N^{-1}h^{-1}\sum_{i=1}^{n}\tilde{X}_i^T K_i(t) Y_i$$

which can be analyzed by the same methods as in Section 4. Because of the limitation of space, we will not pursue such an analysis here. Similarly, smoothing splines and locally weighted polynomials can be obtained for (5.1) by estimating $E\left(XY(t)\right)$ using the corresponding methods.

5.2 Binary Outcomes

The estimation methods of Section 2 are most appropriate for models with continuous outcomes $Y(t)$. In situations such as evaluating the conditional probability of presence or absence of a certain condition, for example a disease, given time t and a set of covariates, $Y(t)$ is a binary variable, i.e.

$$Y(t)=\begin{cases} 1 & \text{if the disease is present at time } t,\\ 0 & \text{otherwise.}\end{cases}$$

Furthermore, conditioning on $X(t)$, $Y(t)$ has a Bernoulli distribution with

$$\mu(t)=P\left(Y(t)=1|X(t)\right).$$

Let Link$(\cdot)$ be a given link function. Generalized time-varying coefficient models are defined by

$$\text{Link}(\mu(t))=X^T(t)\,\beta(t). \tag{5.3}$$

Popular choices of link functions including the logit, the complementary log-log, etc. are discussed, for example, in McCullagh and Nelder (1989).

When the coefficients $\beta=(\beta_0,\ldots,\beta_k)^T$ are independent of t, (5.3) reduces to the well-known generalized linear models. Statistical inferences and estimation methods for this class of models with longitudinal data can be found in Diggle, Liang and Zeger (1994). A penalized likelihood approach (Green and Silverman (1994)) could be taken to allow for time varying coefficients with longitudinal data.

5.3 Estimation at the Boundary

The asymptotic properties developed in Section 4 were built on a crucial assumption of t being an interior point of the observation interval. In practice, it is also important to estimate the $\beta(t)$ values when t is at the boundary. For the example of Section 3, a meaningful goal of model (1.2) is to use the covariates such as the maternal vitamin A level and infant's gender and HIV status to predict infant's weight at birth. It is well known that kernel estimates need special boundary modifications in order to avoid severe bias at the boundary—see, for example, Gasser and Müller (1979), Rice (1984), and Müller (1984). Smoothing splines, too, suffer from increased bias near the boundary because they are constrained to have vanishing second derivatives there (Rice and Rosenblatt (1983)). Locally weighted polynomial smoothings typically have increased variance near the boundary. We have not investigated the boundary behavior of our smoothing procedures for longitudinal data, but we would expect similar phenomena.

5.4 Global Measures of Errors

Our analysis in Section 4 focused on the asymptotic behavior of kernel estimates at a fixed point t. The asymptotics of global measures could be pursued as well.

Several global risks such as the average mean squared errors (AMSE) and the average predictive squared errors (APSE) have been given in Section 2.5. One common feature about AMSE and APSE as defined in (2.21), (2.23) and (2.24) is that they are all defined conditioning on the design points t_{ij}.

It might be more reasonable to measure the average performance of the estimates on the entire interval according to some weight functions $\pi(\cdot)$. Some possible measures are

$$\text{MISE}\left(\hat{\beta}_l\right) = \int \text{MSE}\left(\hat{\beta}_l(t)\right) \pi(t)\, dt, \qquad l = 0, \ldots, k, \tag{5.4}$$

and

$$\text{MISE}_{\mathbf{w}}\left(\hat{\beta}\right) = \int \text{MSE}_{\mathbf{w}}\left(\hat{\beta}(t)\right) \pi(t)\, dt \tag{5.5}$$

where $\text{MSE}\left(\hat{\beta}_l(t)\right)$ is defined in (2.20) and $\text{MSE}_{\mathbf{w}}\left(\hat{\beta}(t)\right)$ is defined in (4.1). To measure the global predictive error of $\hat{\beta}(t)$ for a given covariate $x = (x_0, \ldots, x_k)^T$ which may be functions of t, a natural quantity is the mean integrated predictive squared error (MIPSE)

$$\text{MIPSE}\left(x^T\hat{\beta}\right) = \int \text{MSE}\left(x^T\hat{\beta}(t)\right) \pi(t)\, dt \tag{5.6}$$

where $\text{MSE}\left(x^T\hat{\beta}(t)\right)$ is defined in (2.22).

5.5 Inference

We have used bootstrap standard errors to assess variability, but we have not addressed some other important inferential issues. Various types of confidence regions might be desired: for example, intervals for components or linear combinations of components of $\beta(t)$ for fixed t and simultaneous confidence bands for all t in an interval. Various hypothesis testing problems are of interest as well: for example, a test that a certain component of $\beta(\cdot)$ is identically zero or constant. The bootstrap provides a natural approach to such problems, but the theoretical and practical aspects would require substantial development.

5.6 Data Analysis

There is an extensive methodology of data analysis in linear models, parallels of which should be developed for the current context. For example, it is desirable to identify influential cases. Residuals for longitudinal data are not only individual points but the curves corresponding to individual subjects. Rice and Silverman (1991) use principal components to characterize temporal variation and individual residual curves, but their methods are not directly applicable to the current context.

6 Proofs

Before proving Theorem 1 and the corollaries of Section 4, we first state a useful technical lemma.

LEMMA 1. *Let $a_{(i,j,l_1,l_2)}(t)$ be a function of $(X_{ijl_1}, X_{ijl_2}, t_{ij}, t, \epsilon_i)$ such that, for some positive constants a and b.*

$$\left| E\left(a^2_{(i,j,l_1,l_2)}(t)|t_{ij} = s\right)\right| \le a|s-t|^b. \tag{6.1}$$

Suppose that t_{ij}, $i = 1,\ldots,n$ and $j = 1,\ldots,n_i$, are independent random variables with density f_T, assumptions A1, A4, A5 and A6 are satisfied, and

$$Z_{l_1l_2}(t) = N^{-1}h^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n_i} a_{(i,j,l_1,l_2)}(t)\, K\left(\frac{t-t_{ij}}{h}\right).$$

Then, for all $l_1, l_2, r_1, r_2 = 0, \ldots, k$,

$$\begin{aligned} E(Z_{l_1l_2}(t)Z_{r_1r_2}(t)) \;=\; & \left(1-\frac{\sum_{i=1}^{n} n_i^2}{N^2}\right)\left(\int \phi^{(1)}_{l_1l_2}(t-hu)f_T(t-hu)K(u)\,du\right) \\ & \times\left(\int \phi^{(1)}_{r_1r_2}(t-hu)f_T(t-hu)K(u)\,du\right) \\ & + N^{-2}\left(\sum_{i=1}^{n} n_i^2 - N\right) f_T^2(t)\rho_{(l_1,l_2,r_1,r_2)}(t) \\ & + N^{-1}h^{-1}\phi^{(2)}_{(l_1,l_2,r_1,r_2)}(t)f_T(t)\int K^2(u)\,du \\ & + o\left(N^{-1}h^{-1}\right) + o\left(N^{-2}\left(\sum_{i=1}^{n} n_i^2 - N\right)\right) \end{aligned} \tag{6.2}$$

where

$$\begin{aligned} \phi^{(1)}_{l_1l_2}(s) &= E\left[a_{(i,j,l_1,l_2)}(t)|t_{ij}=s\right], \\ \phi^{(2)}_{(l_1,l_2,r_1,r_2)}(s) &= E\left[a_{(i,j,l_1,l_2)}(t)\, a_{(i,j,r_1,r_2)}(t)|t_{ij}=s\right], \\ \rho_{(l_1,l_2,r_1,r_2)}(s) &= \lim_{\Delta\to 0} E\left[a_{i,j,l_1,l_2}(t)\, a_{i,j',l_1,l_2}(t)|t_{ij}=s,\, t_{ij'}=s+\Delta\right]. \;\square \end{aligned}$$

PROOF: By straightforward decomposition, we have

$$Z_{l_1l_2}(t)Z_{r_1r_2}(t) = A_1 + A_2 + A_3$$

where

$$
\begin{aligned}
A_1 &= N^{-2}h^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n_i} a_{(i,j,l_1,l_2)}(t)a_{(i,j,r_1,r_2)}(t)\,K^2\left(\frac{t-t_{ij}}{h}\right)\\
A_2 &= N^{-2}h^{-2}\sum_{i=1}^{n}\sum_{j\neq j'} a_{(i,j,l_1,l_2)}(t)a_{(i,j',r_1,r_2)}(t)\,K\left(\frac{t-t_{ij}}{h}\right)K\left(\frac{t-t_{ij'}}{h}\right)\\
A_3 &= N^{-2}h^{-2}\sum_{i\neq i'}^{n}\sum_{j,j'} a_{(i,j,l_1,l_2)}(t)a_{(i',j',r_1,r_2)}(t)\,K\left(\frac{t-t_{ij}}{h}\right)K\left(\frac{t-t_{i'j'}}{h}\right).
\end{aligned}
$$

By (6.1), A4 through A6 and the change of variables, it is easy to verify that

$$EA_1 = N^{-1}h^{-1}\phi^{(2)}_{l_1l_2}(t)f_T(t)\int K^2(u)\,du + o\left(N^{-1}h^{-1}\right).$$

Similarly, (6.1), A5 and the Cauchy-Schwarz inequality imply that, for any $j\neq j'$, $u_1\in R$ and $u_2\in R$,

$$E\left(a_{(i,j,l_1,l_2)}(t)a_{(i,j',r_1,r_2)}(t)|t_{ij}=t-hu_1, t_{ij'}=t-hu_2\right)\to \rho_{(l_1,l_2,r_1,r_2)}(t)$$

as $h\to 0$. Thus

$$EA_2 = N^{-2}\left(\sum_{i=1}^{n} n_i^2 - N\right) f_T^2(t)\rho_{(l_1,l_2,r_1,r_2)}(t) + o\left(N^{-2}\left(\sum_{i=1}^{n} n_i^2 - N\right)\right).$$

Finally, direct calculation shows that

$$
\begin{aligned}
EA_3 &= \left(1-\frac{\sum_{i=1}^n n_i^2}{N^2}\right)\left(\int \phi^{(1)}_{l_1l_2}(t-hu)f_T(t-hu)K(u)\,du\right)\\
&\qquad\times\left(\int \phi^{(1)}_{r_1r_2}(t-hu)f_T(t-hu)K(u)\,du\right)
\end{aligned}
$$

The proof of the lemma is completed. □

Proof of Theorem 1: Following (4.3) and (4.4), it suffices to study the asymptotic properties of $E\hat{R}(t)$ and $E\left[\hat{R}_l(t)\hat{R}_r(t)\right]$ for all $l, r = 0,\ldots,k$.

Select $a_{(i,j,l_1,l_2)}(t)$ of Lemma 1 to be

$$a_{(i,j,l_1,l_2)}(t) = \xi_{ijl_1}(t) = \sum_{s=0}^{k}\left[X_{ijl_1}X_{ijs}\left(\beta_s(t_{ij})-\beta_s(t)\right)\right] + X_{ijl_1}\epsilon_i(t_{ij}).$$

Then, it can be verified by direct computation that

$$\hat{R}_l(t) = N^{-1}h^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\xi_{ijl}(t)\,K\left(\frac{t-t_{ij}}{h}\right),\quad l=0,\ldots,k, \tag{6.3}$$

$$(6.4) \qquad E\left(\xi_{ijl}(t)|t_{ij}=s\right)=\sum_{r=0}^{k}\left(\beta_r(t)-\beta_r(t)\right)\rho_{lr}(s),$$

and

$$(6.5) \qquad \begin{aligned} E\hat{R}(t) &= N^{-1}h^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\int E\left(\xi_{ijl}(t)|t_{ij}=s\right)K\left(\frac{t-s}{h}\right)f_T(s)\,ds \\ &= \left(B_0(t),\ldots,B_k(t)\right)^T. \end{aligned}$$

Thus, (4.6) follows from (6.3), (6.4) and (6.5).

For the derivation of (4.7), we first notice that

$$(6.6) \qquad \begin{aligned} & E\left(\xi_{ijl_1}(t)\xi_{ijl_2}(t)|t_{ij}=s\right) \\ &= E\left\{\left[\sum_{r=0}^{k}X_{ijl_1}X_{ijr}\left(\beta_r(t_{ij})-\beta_r(t)\right)\right]\left[\sum_{r=0}^{k}X_{ijl_2}X_{ijr}\left(\beta_r(t_{ij})-\beta_r(t)\right)\right]|t_{ij}=s\right\} \\ &\quad +E\left[X_{ijl_1}X_{ijl_2}\epsilon_i^2(t_{ij})|t_{ij}=s\right] \\ &= \sum_{r_1,r_2=0}^{k}\left(\beta_{r_1}(s)-\beta_{r_1}(t)\right)\left(\beta_{r_2}(s)-\beta_{r_2}(t)\right)E\left[X_{ijl_1}X_{ijl_2}X_{ijr_1}X_{ijr_2}|t_{ij}=s\right] \\ &\quad +\sigma^2(s)\,\rho_{l_1l_2}(s). \end{aligned}$$

Thus, (4.7) follows directly from (6.2), (6.4), (6.5) and (6.6).

By (4.7) and assumptions A2 through A6, it is easy to see that $\mathrm{I}_N(t)$ and $\mathrm{III}_N(t)$ converge to zero as $n\to\infty$. Thus, by the definition of $\mathrm{II}_N(t)$ and the fact that $\sum_{l_1,l_2=0}^{k}M_{l_1l_2}(t)\rho_{l_1l_2}(t)>0$, it suffices to show the last assertion of the theorem by proving that

$$\lim_{n\to\infty}N^{-2}\left(\sum_{i=1}^{n}n_i^2-N\right)=0 \qquad \text{if and only if} \qquad \lim_{n\to\infty}\max_{1\le i\le n}(n_i/N)=0.$$

It is easy to see that $\lim_{n\to\infty}\max_{1\le i\le n}(n_i/N)\neq 0$ implies that $\lim_{n\to\infty}\sum_{i=1}^{n}(n_i/N)^2\neq 0$. It suffices to show that $\lim_{n\to\infty}\max_{1\le i\le n}(n_i/N)=0$ implies $\lim_{n\to\infty}\sum_{i=1}^{n}(n_i/N)^2=0$.

Assume now that $\lim_{n\to\infty}\max_{1\le i\le n}(n_i/N)=0$. Then, for any $\epsilon>0$, $\max_{1\le i\le n}(n_i/N)<\epsilon/2$ for sufficiently large n. Let $1=k_0<k_1<\cdots<k_m=n_n$ be positive integers such that

$$\frac{\epsilon}{2}<\sum_{i=k_{l-1}}^{k_l}\frac{n_i}{N}<\epsilon \qquad \text{for } l=1,\ldots,m-1, \text{ and } \sum_{i=k_{m-1}}^{k_m}\frac{n_i}{N}<\epsilon.$$

Then, for all $l=1,\ldots,m$, $\sum_{i=k_{l-1}}^{k_l}(n_i/N)^2<\epsilon^2$. By $N=\sum_{i=1}^{n}n_i$, we have $m\le 2/\epsilon$, and consequently,

$$\sum_{i=1}^{n}\left(\frac{n_i}{N}\right)^2<\frac{2\epsilon^2}{\epsilon}=2\epsilon.$$

Since ϵ can be arbitrarily small, the above inequality implies that $\lim_{n\to\infty} \sum_{i=1}^{n}(n_i/N)^2 = 0$. $\square$

PROOF OF COROLLARY 1: By A 5 and A6, (4.9) is a direct consequence of (4.6), the definition of dth order kernels and the Taylor expansions of $\beta_r(t-hu)$, $\rho_{lr}(t-hu)$ and $f_T(t-hu)$ at point t. $\square$

PROOF OF COROLLARY 2: By (4.10), A1 and A2, there exists a constant $b>0$ so that

$$\left|M_{l_1 l_2}(t)\sigma^2(t)E\left(X_{ijl_1}X_{ij'l_2}|t_{ij}=t_{ij'}=t\right)\right| \le b$$

for all $t \in R$. On the other hand, (4.11) implies that

$$N^{-2}\sum_{i=1}^{n}\sum_{j\neq j'} b = bN^{-2}\left(\sum_{i=1}^{n} n_i^2 - N\right) \le b\lambda N^{-1} - bN^{-1} = o\left(N^{-1}h^{-1}\right)$$

when n is sufficiently large. Thus (4.12) follows from (4.10). $\square$

PROOF OF COROLLARY 3: Denote by $A(h,t)$ the sum of the first two terms of the right hand side of (4.12). Then $A(h,t)$ is minimized by h_{opt} which is the unique solution of $\partial A(h,t)/\partial h = 0$. It is easy to verify that (4.14) is obtained by substituting h_{opt} into (4.12). $\square$

Acknowledgments. This research was partially supported by grants from the National Institutes of Health, AI26499, AI33874 and HD30042, and the Office of Nutrition, Bureau for Science and Technology, United States Agency for International Development Cooperative Agreement DAN-0045-A-5094-00. The authors are grateful to Professor Richard D. Semba for providing us his valuable data on children's growth and vitamin A, to Professor Shih-Ping Han for insightful discussions and comments on the properties of splines, and to Mr. Chin-Tsang Chiang for many numerical computations used in this paper.

REFERENCES

ALTMAN, N. S. (1990). Kernel smoothing of data with correlated errors. *J. Amer. Statist. Assoc.* **85**, 749-759.

BUJA, A., HASTIE, T. J. and TIBSHRIANI, R. J. (1989). Linear smoothers and additive models. *Ann. Statist.* **17**, 453-555.

CLEVELAND, W. S. (1979). Robust locally-weighted regression and smoothing scatterplots. *J. Amer. Statist. Assoc.* **74**, 829-836.

COOK, R. D and WEISBERG, S. (1982). *Residuals and Influence in Regression.* Chapman and Hall, New York.

DIGGLE, P. J. (1988). An approach to the analysis of repeated measurements. *Biometrics* **44**, 959-971.

DIGGLE, P. J., LIANG, K. Y. and ZEGER, S. L. (1994). *Analysis of Longitudinal Data.* Oxford University Press, Oxford, England.

EUBANK, R. L. (1988). *Spline Smoothing and Nonparametric Regression.* Marcel Dekker, New York.

FAN, J. Q. (1993). Local linear regression smoothers and their minimax efficiencies. *Ann. Statist.* **21**, 196-216.

FAN, J. Q. and MARRON, J. S. (1992). Best possible constant for bandwidth selection. *Ann. Statist.* **20** 2057-2070.

GASSER, T. and MÜLLER, H. G. (1979). Kernel estimation of regression functions. In: *Smoothing Techniques for Curve Estimation,* eds. Gasser and Rosenblatt. Springer-Verlag, Heidelberg.

GREEN, P. J. and SILVERMAN, B. W. (1994). *Nonparametric Regression and Generalized Linear Models.* Chapman and Hall, London.

HALL, P., SHEATHER, S. J., JONES, M. C. and MARRON, J. S. (1991). On optimal data-based bandwidth selection in kernel density estimation. *Biometrika* **78** 263-269.

HÄRDLE, W. (1990). *Applied Nonparametric Regression.* Cambridge University Press, Cambridge, England.

HÄRDLE, W. and MARRON, J. S. (1983). The nonexistence of moments of some kernel regression estimators. *North Carolina Institute of Statistics, Mimeo Series* No. 1537.

HART, J. D. (1991). Kernel regression estimation with time series errors. *J. Roy. Statist. Soc. Ser. B* **53**, 173-187.

HART, J. D. and WEHRLY, T. E. (1986). Kernel regression estimation using repeated measurements data. *J. Amer. Statist. Assoc.* **81**, 1080-1088.

HASTIE, T. J. and LOADER, C. (1993). Local regression: automatic kernel carpentry. *Statist. Sci.* **8** 120-143.

HASTIE, T. J. and TIBSHRIANI, R. J. (1990). *Generalized Additive Models.* Chapman and Hall, London.

HASTIE, T. J. and TIBSHRIANI, R. J. (1993). Varying-coefficient models. *J. Roy. Statist. Soc. B* **55**, 757-796.

JONES, R. H. (1987). Serial correlation in unbalanced mixed models. *Bull. Internat. Statist. Inst.* **46**, 105-122.

JONES, R. H. and ACKERSON, L. M. (1990). Serial correlation in unequal spaced longitudinal data. *Biometrika* **77**, 721-731.

JONES, R. H. and BOADI-BOTENG, F. (1991). Unequally spaced longitudinal data with serial correlation. *Biometrics* **47**, 161-175.

MARRON, J. S. and WAND M. P. (1992). Exact mean integrated squared error. *Ann. Statist.* **20**, 712-736.

MCCULLAGH, P. and NELDER, J. A. (1989). *Generalized Linear Models.* 2nd Edition. Chapman and Hall, London.

MOYEED, R. A. and DIGGLE, P. J. (1994). Rates of convergence in semi-parametric modeling of longitudinal data. *Austral. J. Statist.* 36, 75-93.

MÜLLER, H. G. (1984). Boundary effects in nonparametric curve estimation models. In: *COMPSTAT*, 84-89, Physica, Verlag.

PANTULA, S. G. and POLLOCK, K. H. (1985). Nested analysis of variance with autocorrelated errors. *Biometrics* **41**, 909-920.

PARKER, R.L, and RICE, J.(1984). Comment on "Some aspects of the spline smoothing approach to nonparametric regression curve estimation," by B.W. Silverman. *Journal of the Royal Statistical Society B*, 1-52.

RICE, J. and ROSENBLATT, M. (1983). Smoothing splines: regression, derivatives, and deconvolution. *Annals of Statistics* **11**, 141-56.

RICE, J. A. (1984). Boundary modification for kernel regression. *Commun. in Statist., Ser. A* **13**, 893-900.

Rice, J. A. and Silverman, B. W. (1991). Estimating the mean and covariance structure nonparametrically when the data are curves. *J. Roy. Statist. Soc. Ser. B* **53**, 233-243.

Rosenblatt, M. (1969). Conditional probability density and regression estimators. *Multivariate Analysis* **2** (P.R. Krishnaiah, ed.) 25-31, North-Holland Pub. Co., Amsterdam.

Semba, R. D. (1994). Vitamin A immunity and infection. *Clin. Inf. Dis.* **19**, 489-499.

Semba, R. D., Miotti, P., Chiphangwi, J. D., Henderson, R., Dallabetta, G., Yang, L. P. and Hoover, D. R. (1996). Maternal vitamin A deficiency and child growth failure during human immunodeficiency virus infection. *Unpublished manuscript.*

Stone, C. J. (1977). Consistent nonparametric regression. *Ann. Statist.* **5**, 549-645.

Ware, J. H. (1985). Linear models for the analysis of longitudinal studies. *Amer. Statist.* **39**, 95-101.

Zeger, S. L. and Diggle, P. J. (1994). Semiparametric models for longitudinal data with application to CD4 cell numbers in HIV seroconverters. *Biometrics* **50**, 689-699.